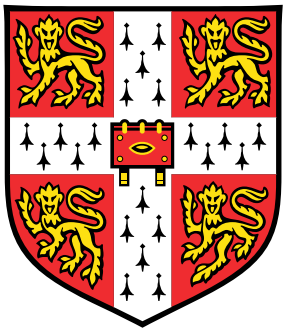
Quantitative super-resolution imaging of cell polarity proteins using DNA-PAINT



Edo Džafić

Department of Genetics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Peterhouse October 2019

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except when specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution. It does not exceed the prescribed word limit for the relevant Degree Committee.

Edo Džafić 27th of March 2020

Quantitative super-resolution imaging of cell polarity proteins using DNA-PAINT

Edo Džafić

Abstract

Knowing the localisation and spatial organisation of proteins is crucial for understanding their function. The development of super-resolution imaging has improved our ability to garner this information, but counting individual molecules in densely-packed assemblies is still challenging. DNA-based point accumulation for imaging in nanoscale topography (DNA-PAINT) is one of the most recently developed imaging techniques in super-resolution microscopy. It uses fluorescently-labelled DNA to visualise the molecules of interest with nanometre precision. DNA-PAINT was initially reliant on antibody labelling of *in vitro* protein targets, however, there is need for an alternative labelling strategy as good antibodies do not exist for many target proteins. Moreover, it is impossible to quantify antibody labelling efficiency, which is a crucial parameter for quantitative imaging. In order to address these issues, I present here an optimised imaging pipeline for protein counting in a thick tissue sample, tens of microns away from the coverslip, for which cell polarity proteins in epithelial cells of the fruit fly (*Drosophila melanogaster*) egg chambers are given as an example.

Firstly, I established an alternative labelling strategy to label polarity proteins for DNA-PAINT imaging using genetically-encoded Halo and SNAP self-labelling enzymes in fruit fly tissue. In this approach, the Halo and SNAP ligands conjugated to DNA react with their respective enzymes to form a covalent bond with the protein of interest in a 1:1 stoichiometry. I then optimised the labelling protocol for imaging the fixed fruit fly tissue and analysed non-specific signal to reduce background during image post-processing. A quantitative Western blot-based gel band shift assay was developed to determine the labelling efficiency of target proteins. Moreover, I used nucleoporin proteins in the nuclear pore complex to calibrate the influx rate of fluorescently-labelled DNA to quantify the number of molecules in super-resolution images. Additionally, I used nucleoporin-160 and nucleoporin-188 to benchmark two-colour super-resolution imaging using DNA-PAINT.

Super-resolution imaging of three apical polarity proteins (aPKC, Crumbs, Par6) in the fruit fly egg chambers revealed that they form mesoscopic-sized clusters along the cell junctions. In order to analyse these clusters in a quantitative manner, I collaborated with

Leila Muresan to develop an image analysis pipeline. My analysis demonstrated that apical polarity proteins are less concentrated in the cytosol by approximately one order of magnitude. To expand on these observations, the junctional clusters were identified by a mean-shift algorithm and classified according to size, i.e. the number of molecules. The cluster size distribution was then approximated by a mathematical function. The model selection was performed by Bayesian information criteria that was tested on simulated data beforehand.

This work provides an optimised imaging pipeline for quantifying the number of protein molecules in a thick biological sample using DNA-PAINT, and proposes a post-processing approach to identify and mathematically describe molecular clustering. These data will prove useful for modelling the spatial organisation of polarity proteins, and provide a framework for greater insight into the biological function of individual proteins.

Acknowledgements

This thesis would not have been created without direct or indirect help from multiple people. I would like to first thank Daniel St Johnston for giving me the opportunity to perform my research in his laboratory and the Wellcome Trust for funding my studies. My immense gratitude goes to George Sirinakis and Edward Allgeyer for continuous support with the technical aspect of imaging, constructive criticism of my experimental work and comments on my thesis chapters. Similarly, I cannot thank enough Leila Muresan for all the help with the computational part of this thesis and comments on one of my thesis chapters. An enormous thank goes to my postdoc mentor David Jordan for continuous help with the mathematical aspect of this thesis and for answering to all my Matlab-related questions.

I would like to thank all the current and previous members of St Johnston lab for help and suggestions with my experiments. Especially, I would like to thank Avik Mukherjee for introducing me to the project at the beginning. An enormous thank to Nick Lowe who made most of the fly lines that I was using in my research. I'm also grateful to Amandine Palandri and Jenny Richens for making the fly lines with endogenously-tagged nucleoporins, which were crucial for quantifications performed in this thesis. I would like to thank Mihoko Tame for giving me comments on my first chapter and support all along, and to Andrew Plygawko for correcting my devastating grammar mistakes. I would like to thank Richard Butler and Nicola Lawrence from the Gurdon Institute's Imaging Facility for all the help with the image analysis and Airy-scan imaging, respectively. I would also like to thank Kathryn Lilley for initial mass spectrometry analysis of labelling efficiency, despite not using this approach in the end, and Frédéric Daste for advice with the sample preparation. Thank you Pierre-François Lenne for initial inspiration and discussions about cluster analysis. Thank you Julia Falo Sanjuan for all Matlab-related support. I should not forget to mention Caren Norden that has been of a great support even after I moved to Cambridge. I would like to thank my tutor Christopher Lester and Peterhouse college for the financial support of my extracurricular activities. Furthermore, I would also like to acknowledge Ewa Paluch and Josana Rodriguez for critical reading of my thesis.

Finally, I would like to thank my friends, you know who you are, each of you contributed in a mosaic of support and I'm extremely grateful for that. I would also like to thank

my parents and my brother for continuous support through all stages of my education. Finally, finishing and writing up this thesis would not have been as nice if it were not for Bruno. I dedicate this thesis to him.

Abbrevations

ALE	absolute labelling efficiency
aPKC	atypical protein kinase C
BG	benzylguanine
BSA	bovine serum albumine
CA	chloroalkane
CAAX	farnesylation signalling sequence; membrane targeting motif
CRIB	Cdc42- and Rac-interactive binding domain
DBSCAN	density-based spatial clustering of applications with noise
Dlg	Discs large
DNA-PAINT	DNA point accumulation for imaging in nanoscale topography
dSTORM	direct stochastic optical reconstruction
DTT	dithiothreitol
ECM	extracellular matrix
EDTA	ethylenediamine tetraacetic acid
EGF	epidermal growth factor
ELE	effective labelling efficiency
EM	electron multiplying
EMCCD	electron multiplying charge-coupled device
F-actin	filamentous actin
FERM	4.1R, ezrin, radixin, moesin binding domain

FLP flippase

FRAP fluorescence recovery after photobleaching

FRET fluorescence resonance energy transfer

FRT flippase recognition target

FWHM full width at half maximum

G-actin globular actin

GAP GTPase-activating protein

GBP GFP binding protein

GFP green fluorescent protein

GPI glycophosphatidylinositol

ISA integrated signal anisotropy

kDa nkilo Dalton

kU kilo enzyme unit

kW kilo Watt

LAT linker for activation of T cells

lbFCS localization-based fluorescence correlation spectroscopy

Lgl lethal giant larvae

MAGUK membrane-associated guanylate kinase

MOPS 3-(N-morpholino)propanesulfonic acid

NIR near-infrared

NPC nuclear pore complex

Nup nucleoporin

PALM photoactivated localization microscopy

Par3 partitioning defective protein 3

Par6 partitioning defective protein 6

PB1 Phox and Bem1 domain

PBS phosphate buffered saline

PCF pair correlation function

PDZ a postsynaptic density-95, discs large, zonula occludens1 binding domain

PFA paraformaldehyde

PKC3 protein kinase C-like 3

PMT photomultiplier tube

PSF point spread function

PTEN phosphatase and tensin homolog

qPAINT quantitative DNA-PAINT

RIPA radioimmunoprecipitation assay buffer

ROI region of interest

sCMOS scientific Complementary metal–oxide–semiconductor

SDS sodium dodecyl sulfate

SMLM single molecule localisation microscopy

TBS Tris-buffered saline

TIRF total internal reflection fluorescence

TOPO cloning with DNA topoisomerase

Tris tris(hydroxymethyl)aminomethane

UAS upstream activating sequence

Contents

Chapter 1

Introduction

1.1 Biological polarity across scales

In biology, polarity is defined by the asymmetric spatial distribution of the components of the biological unit. Biological polarity exists across different scales from molecules to organisms.

One of the most stereotypical examples of polarity at the molecular level is the globular actin (G-actin) monomer. These monomers can polymerise into filamentous actin (F-actin) that is also a polarized structure. Similarly, α- and β-tubulin are polarized monomers that assemble into microtubules. Both these cytoskeletal structures support cell polarity. One of the most striking examples of it are neuronal and epithelial cells. In the case of neuronal cells, long membrane projections called axons transduce the electrical signal, while in epithelial cells the plasma membrane facing the lumen is ruffled into microvilli for uptake of molecules. Polarized cells in turn build polarized tissues and organs (e.g. the intestine). Finally, multicellular organisms exhibit body plan polarity as well, often with multiple axes (e.g. head-tail, dorsal-ventral, left-right). Loss of polarity often results in tissue disintegration and developmental defects (McCaffrey and Macara 2009).

One could of course continue this ascent across scales, however this is not the main goal of this opening paragraph. However, it is important to grasp that polarity facilitates functional complexity such as differentiation (Lee et al. 2006), growth (Bilder et al. 2000), motility (Nishio et al. 2007) and pattern formation (Goehring and Grill 2013). This is probably best understood at the cellular level. Almost all cells studied so far exhibit some degree of polarity. Cell polarity is reflected as an asymmetric distribution of membrane

lipids and membrane-associated proteins along an axis.

Because the work presented in this thesis focuses on polarity proteins in epithelial cells, I will first give an overview of the epithelial cells and polarity proteins. Next, I will present our current understanding of cell polarity regulation in different systems and in particular in the fruit fly (*Drosophila melanogaster*) and the worm (*Caenorhabditis elegans*). Finally, I will explain how single molecule imaging in the worm embryo has brought a novel understanding of cell polarity regulation.

1.2 Epithelial cell characteristics

Epithelial cells are one of the four basic cell-types in multicellular animals. Their plasma membrane is asymmetrically organised into three primary domains: apical, lateral and basal (Figure 1.1) (Simons and Fuller 1985). The apical domain usually faces the lumen (external environment) and is enriched with filamentous actin (F-actin) that supports the formation of microvilli. Below the apical domain there is a subapical domain (here referred to as the marginal zone). At the boundary between the marginal zone and the lateral membrane are the adherens junctions (here referred to as a domain as well) that hold cells together through transmembrane proteins called cadherins (Harris and Tepass 2010). Additionally, at the border between apical and lateral domain there are sealing junctions (Zihni et al. 2016). In vertebrates, these are called tight junctions and are positioned above adherens junctions. In invertebrates, they are called septate junctions and are positioned below adherens junctions. While the lateral and basal membranes are contiguous (and therefore in the literature usually referred as basolateral), only the basal membrane is in contact with the extracellular matrix (ECM) through transmembrane proteins called integrins (Lee and Streuli 2014).

1.3 Overview of polarity proteins

In the last few decades cell polarity genes have been identified and their proteins characterised across different cell types and organisms. Throughout the literature polarity proteins have been classified either based on their biochemical nature (kinases vs scaffold proteins) or subcellular localisation (apical vs lateral, anterior vs posterior). I will use the latter classification, because this work primarily focuses on the apical domain. Therefore,

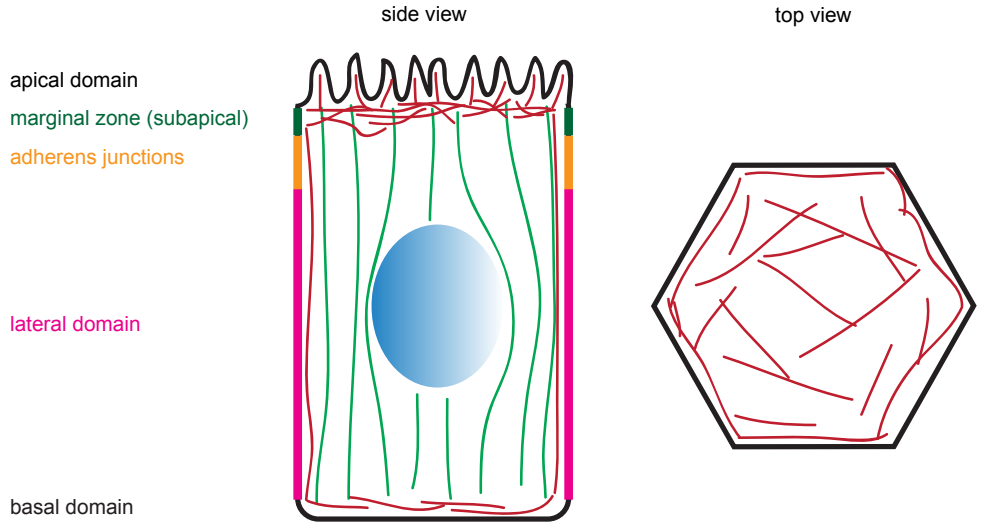


Figure 1.1 **A schematic view of an epithelial cell.** Left: side view, respective domains are colour-coded. Green lines indicate microtubules, red lines indicate F-actin. Right: top view, cross-section through the marginal zone. Red lines indicate F-actin.

I will describe apical polarity proteins in the next paragraphs in more detail in the context of fruit fly epithelia and then briefly mention lateral proteins as well.

1.3.1 Apical polarity proteins

Crumbs

Crumbs protein is the only transmembrane member among the polarity proteins. Crumbs was identified in *D. melanogaster* screen, where its mutant caused disruption of the embryonic cuticle and thus a crumbs-like appearance (Tepass et al. 1990). The protein consists of a large extracellular domain with epidermal growth factor (EGF)-like repeats and laminin A G-domain-like repeats, a single-span transmembrane domain and a short 37 amino acid long conserved cytoplasmic tail. The intracellular tail contains a FERM (F for 4.1 protein, E for ezrin, R for radixin, and M for moesin)-binding domain and a PDZ (P for postsynaptic density-95, D for discs large, and Z for zonula occludens1)-binding domain (Figure 1.2). The extracellular domain has been implicated in cis and trans Crumbs oligomerization (Letizia et al. 2013; Röper 2012; Zou et al. 2012).

Crumbs' FERM binding domain can bind Yurt (Laprise et al. 2006), moesin and βheavy-spectrin (Médina et al. 2002). The PDZ binding domain can bind to Par6 or to Stardust (PALS1 in mammals) (Hong et al. 2001), a membrane-associated guanylate kinase

(MAGUK). Immunoelectron microscopy demonstrated that Crumbs localises to the membrane apical to adherens junctions, which has been defined as the marginal zone of the apical membrane (Tepass 1996).

Crumbs is expressed in most, but not all, epithelia in the fruit fly. In some epithelia Crumbs appears to be redundant with other apical polarity proteins, for example Par3 (Tanentzapf and Tepass 2003). When not redundant, Crumbs loss causes defects in epithelial tissue integrity (Campbell et al. 2009).

There is no Crumbs homologue in the worm (*C. elegans*). In mammals there are three Crumbs genes. CRB1 encodes a protein that is a *D. melanogaster* homologue. Additionally, there are CRB2 and CRB3. CRB2 is required for early morphogenesis of the mouse embryo (Xiao et al. 2011). CRB3 lacks the extracellular domain (Makarova et al. 2003).

Atypical protein kinase C

Atypical protein kinase C (aPKC) is one of the three subfamilies of serine/threonine PKC enzyme family. The two other subfamilies are classic PKC and novel PKC. The protein consists from the N-terminal Phox and Bem1 (PB1) domain and a serine/threonine kinase domain (Figure 1.2). PB1 domain binds to the PB1 domain of another polarity protein Par6, which consistently co-localizes with aPKC as assessed by fluorescence confocal microscopy or co-immunoprecipitation experiments (Yamanaka et al. 2001).

Like Crumbs, aPKC localises to the marginal zone in epithelial cells, according to fluorescence confocal microscopy. It was thought that its subcellular localisation relies exclusively on protein-protein interactions. Hence it has been assumed that aPKC is mainly localised at the cortex through interaction with Crumbs (Sotillos et al. 2004), Stardust (Wang et al. 2004), Par3 (Morais-de-Sá et al. 2010) or Cdc42 (Joberty et al. 2000; Lin et al. 2000).

However, recently it was demonstrated that the pseudosubstrate region of aPKC acts as a polybasic domain, which is sufficient to target aPKC to the plasma membrane via electrostatic binding to PM phosphoinositides (Dong et al. 2019). Loss of aPKC in the fruit fly embryos results in the failure of polarity maintenance due to perturbed adherens junction positioning via microtubule cytoskeleton (Harris and Peifer 2007).

aPKC has one homologue in the worm called PKC-3. In mammals there are two homologues: aPKC-λ/ι and aPKC-γ. While mammalian homologues seem to be redundant for polarity,

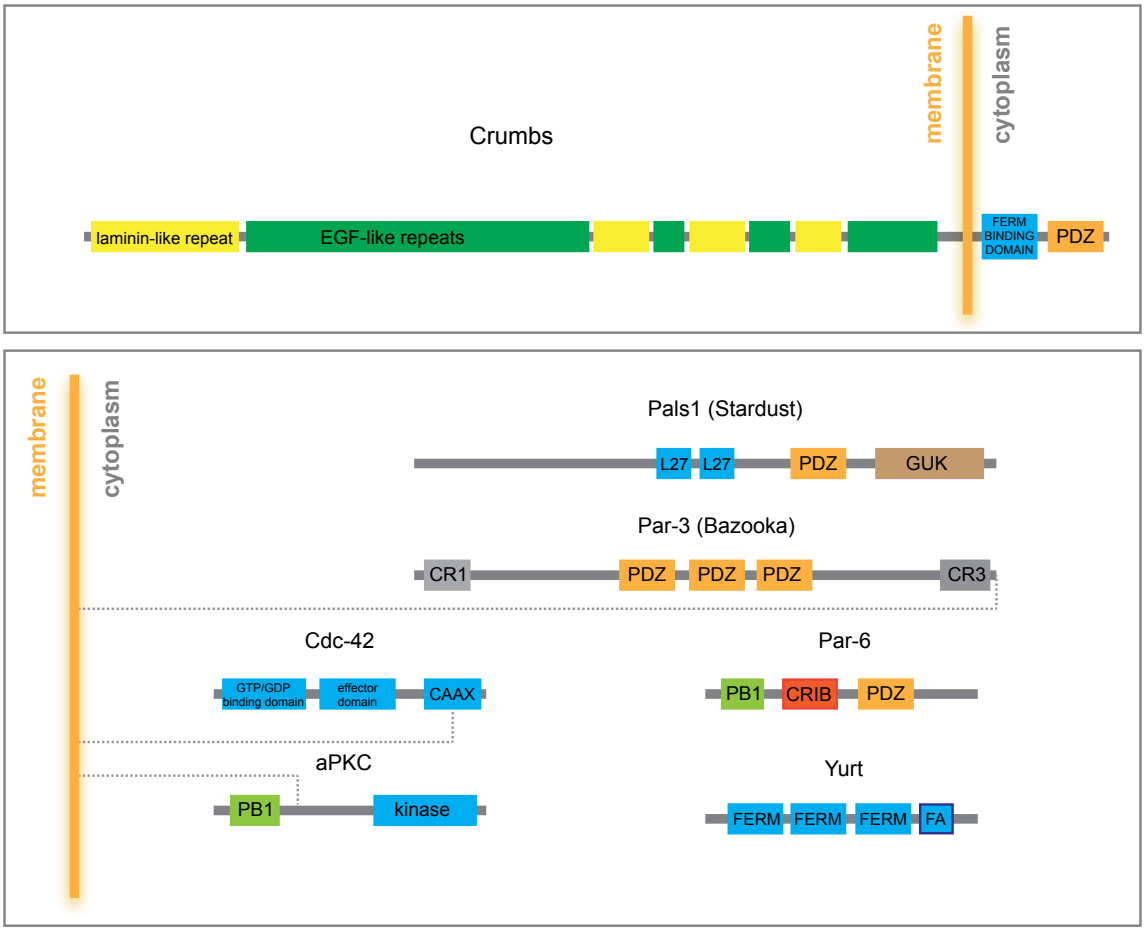


Figure 1.2 **Domain organisation of the apical polarity proteins.** Abbreviations: aPKC, atypical protein kinase C; CRIB, Cdc42/Rac interactive binding domain; EGF, epidermal growth factor; ECR, evolutionary conserved region; FERM, Four-point-one, Ezrin, Radixin, Moesin; FERM-FA, FERM adjacent domain; GUK, guanylate kinase; PB1, Phox and Bem1 domains. Proteins are not drawn in scale. Redrawn from (Tepass 2012).

they seem to have acquired new cellular functions, like stem cell-self renewal (Mah et al. 2015) and intracellular signalling (Hirai and Chida 2003).

Partitioning-defective protein 6

Partitioning-defective protein 6 (Par6) is an adaptor protein that was discovered in a genetic screen for mutations that affect asymmetric cell division in the worm embryo (Kemphues et al. 1988; Tabuse et al. 1998; Watts et al. 1996). It contains a PB1 domain, a Cdc42/Rac interacting binding (CRIB) domain and a PDZ domain (Figure 1.2).

As already mentioned the PB1 domain binds to the PB1 domain of aPKC (Yamanaka et al. 2001). The PDZ domains binds to the PDZ-binding motif of Crumbs (Hong et al.

2001; Nam and Choi 2003). This binding is enhanced after interaction of Cdc42-GTP with the CRIB domain (Whitney et al. 2016).

Par6 localises to the marginal zone, where it is complexed with aPKC, according to fluorescence confocal microscopy. It is enriched cortically and it is thought that it cannot directly bind to plasma membrane and relies exclusively on protein-protein interactions. Recently it was demonstrated that Par6 is necessary for aPKC binding to plasma membrane (Dong et al. 2019).

Loss of Par6 causes similar defects as loss of aPKC (Hutterer et al. 2004; Petronczki and Knoblich 2001; Wodarz et al. 2000). There are three homologues in mammals: PAR6A, PAR6B and PAR6G.

Partitioning-defective protein 3

Like Par6, Partitioning-defective protein 3 (Par3) was discovered in genetic screens in the worm embryo (Etemad-Moghadam et al. 1995). Par3 is called Bazooka in *D. melanogaster*. There are two homologues in mammals: PAR3A and PAR3B. It contains several conserved functional domains, including an N-terminal (CR1) oligomerization domain, three PDZ domains and a C-terminal (CR3) (Figure 1.2).

PDZ domains can bind to β-catenin at adherens junctions but also Par6 and lipid phosphatase PTEN. Moreover it can interact with aPKC, Stardust (PALS1 in mammals), and Par1. Additionally, 14-3-3 proteins, Rho kinase (Nakayama et al. 2008) and phosphoinositides (Stein et al. 2005) have been all demonstrated to interact with Par3.

When interacting with aPKC and Par6, Par3 can be found at the marginal zone. However, in polarized cells, Par3 is enriched and localises at the adherens junctions, below the marginal zone. These localisations were inferred from fluorescence confocal microscopy (Morais-de-Sá et al. 2010). Loss of Par3 leads to defects in establishing cell polarity and mitotic spindle orientation (Hao et al. 2010; Huynh et al. 2001). Moreover loss of Par3 promotes tumorigenesis (Xue et al. 2013).

Cdc42

Cdc42 is the small GTPase of the Rho family that was discovered in *S. cerevisiae* (Adams et al. 1990). It contains a GTP/GDP binding domain, an effector domain, a Rho insertion

domain and a CAAX motif (Figure 1.2). Cdc42 is one of the most conserved polarity proteins across the metazoans (Cotteret and Chernoff 2002).

Cdc42 can exist in active (GTP-loaded) or inactive (GDP-loaded) states. In order for Cdc42 to become active, it has to bind a guanine nucleotide exchange factor (GEF). Which GEFs proteins recruit and control Cdc42 in the fruit fly is not known. In sillico modelling in the St Johnston lab to predict Cdc42-GEF interaction identified several candidates, however they have not yet been characterized (Avik Mukherjee, PhD thesis). Active Cdc42 binds to the CRIB domain of Par6.

Cdc42 is cortically enriched and it is thought that the active form (Cdc42-GTP) is localised at the marginal zone (Fletcher et al. 2012). However, upon prenylation of the CAAX motif, it can also bind to the plasma membrane (Nishimura and Linder 2013).

Overexpression of constitutively active Cdc42 in Drosophila epithelial cells causes ectopic spreading of the apical polarity proteins around the plasma membrane, cell rounding and loss of polarity (Fletcher et al. 2012). On the other hand, expression of dominant-negative Cdc42 causes expansion of the basolateral domain (Genova et al. 2000).

1.3.2 Lateral polarity proteins

The lateral (often referred as basolateral) polarity proteins include Scribble, lethal giant larvae (Lgl) and discs-large (Dlg), also referred as the Scribble complex based on the co-immunoprecipitation and genetic experiments (Bilder et al. 2000). Additionally, the partitioning-defective protein 1 (Par1) kinase localises to the lateral membrane. Par1 phosphorylates Par3 (Benton and St Johnston 2003). Moreover, FERM domain protein Yurt is also considered part of the lateral polarity proteins group because of the important mechanistic role that it has in the mutual antagonism (Laprise et al. 2006).

1.4 Classical view of cell polarity across different model systems

Cell polarity has been traditionally studied in different organisms and cell types. In the next paragraphs I will describe the discoveries that have been made in the worm embryo and in fruit fly epithelia. Based on studies in different experimental systems, a common

model of cell polarity regulation bridging similarities and explaining differences is often proposed (for reviews see (Lang and Munro 2017)). In general, both, polarity establishment and maintenance are thought to depend on mutual antagonism between apical and lateral proteins (Figure 1.3 and 1.4). The reader should note that in general anterior-posterior polarity proteins in the worm embryo are analogous to apical-lateral polarity proteins in the epithelial cells.

1.4.1 In the worm (*C. elegans*) embryo

The worm zygote is initially not polarised. Before polarity establishment, the future anterior polarity proteins PKC3, PAR6, and PAR3 are uniformly distributed throughout the cortex. On the other hand, the future posterior polarity proteins are uniformly cytoplasmic. After the entry of the sperm, its centrosome acts locally to inhibit RhoA-dependent cortical actomyosin contractility (Cowan and Hyman 2004; Bienkowska and Cowan 2012). This asymmetric contractility creates anterior-directed cortical flows that segregate PKC3, PAR6, and PAR3 to the anterior pole (Goehring et al. 2011a; Munro et al. 2004). Microtubules from the sperm centrosome can also promote PAR2 association at the posterior cortex to recruit PAR1, which phosphorylates and promotes dissociation of PAR3 (Motegi et al. 2011; Boyd et al. 1996). Hence, the symmetry breaking can occur also in the absence of the cortical contractility.

After initial polarity has been established, the maintenance phase starts (Figure 1.3). During this phase, active CDC-42 becomes enriched at the anterior pole, where it interacts with anterior polarity proteins and activates them. While at the posterior pole, GTPase-activating protein (GAP) for Cdc42, called CHIN1 becomes highly enriched. Consequently Cdc42 is inactive at the posterior (Kumfer et al. 2010).

1.4.2 In fruit fly (*D. melanogaster*) epithelia

Polarity has been extensively studied in the fruit fly in two different epithelial types: the primary epithelium of the embryo and the follicular epithelium in the egg chambers of adult flies. Importantly, in epithelial cells, polarity proteins segregate into three domains: apical, junctional (adherens junctions), and lateral. Although the principles are analogous to the worm embryo, the molecular mechanism are more complex and less understood (Figure 1.4).

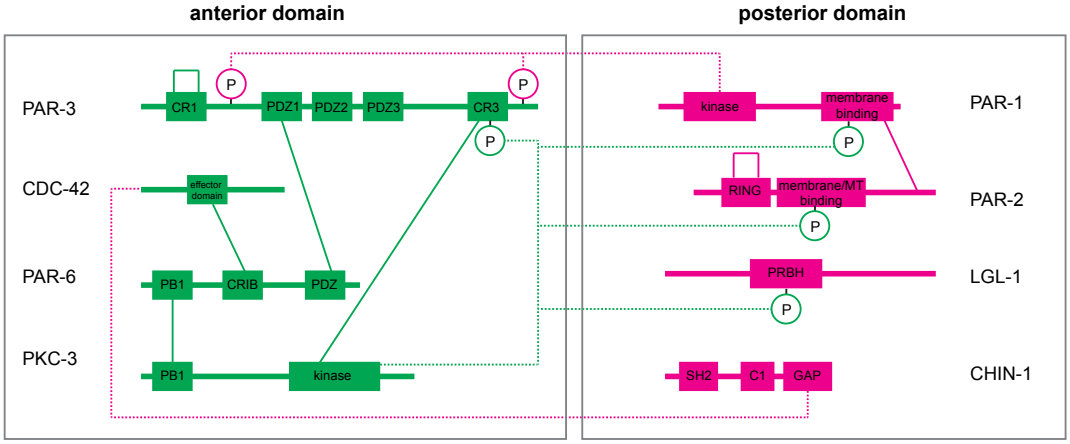


Figure 1.3 **Core molecular interactions between polarity proteins in the worm embryo.** A schematic view of the polarity proteins network indicating key domains and phosphorylation sites involved in protein-protein interaction. Solid lines indicate direct binding interactions. Dotted lines terminating in circles represent phosphorylation. Self-connecting loops indicate oligomerization. Redrawn from (Lang and Munro 2017).

In the embryonic epithelium, polarization occurs simultaneously with cellularization, but importantly the cytoskeleton is already polarized (Schmidt et al. 2018). Par3, acts as the main upstream protein required for establishment of the apical domain through assembly of adherens junctions (Harris and Peifer 2005). On the other hand, it has been thought that in the follicular epithelium, symmetry breaking occurs after contact with ECM and establishment of the basal domain where integrins localise (Schneider et al. 2006; Tanentzapf et al. 2000), however recent re-evaluation of early work failed to confirm this (Lovegrove et al. 2019). Similarly, in the follicular epithelium, Par3 is also thought to be one of the most upstream polarity proteins to initiate establishment of the apical domain (Benton and St Johnston 2003; Franz and Riechmann 2010; Morais-de-Sá et al. 2010), however there is one study claiming that Par3 is dispensable (Shahab et al. 2015).

During polarity establishment Par3 transiently interacts with Par6-aPKC, most likely to facilitate their association with the apical membrane (Achilleos et al. 2010; Franz and Riechmann 2010; Harris and Peifer 2005; Horikoshi et al. 2009; Krahn et al. 2010; McCaffrey and Macara 2009; McKinley and Harris 2012). This transient interaction has been thought to be possible because CR3 domain of Par3 inhibits aPKC kinase activity and traps Par6-aPKC complex in a high-affinity state (Soriano et al. 2016). However, recent re-evaluation of these findings by using more sensitive assay failed to reproduce them (Holly and Prehoda 2019; Thompson and McDonald 2019). Hence Par3 does not inhibit aPKC kinase activity but this does not exclude possibility that additional steps

are involved in stimulating aPKC to phosphorylate Par3.

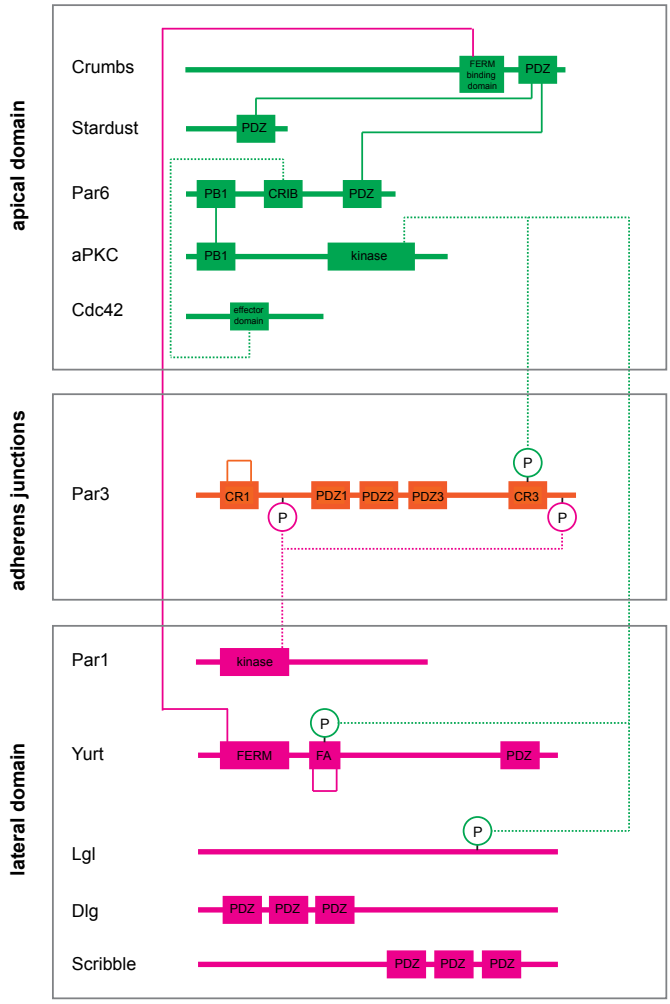


Figure 1.4 **Core molecular interactions between polarity proteins in Drosophila epithelial cells.** A schematic view of the polarity proteins network indicating key domains and phosphorylation sites involved in protein-protein interaction. Solid lines indicate direct binding interactions. Dotted lines terminating in circles represent phosphorylation. Self-connecting loops indicate oligomerization. Inspired by (Lang and Munro 2017).

Nevertheless, phosphorylation of Par3 by aPKC results in binding of Par6-aPKC complex to PDZ domain of Crumbs (Morais-de-Sá et al. 2010; Walther and Pichaud 2010). This binding to Crumbs might be enhanced by binding of Cdc42 to Par6 and increasing its affinity for Crumbs (Whitney et al. 2016). However, it is important to point out that in the context of the follicular epithelium loss of Crumbs does not necessarily result in loss of aPKC (Sherrard and Fehon 2015). As already mentioned, a recent study demonstrated that aPKC complexed with Par6 can bind plasma membrane as well. In this context Par6

inhibits aPKC kinase activity. Interestingly, this membrane-bound fraction of aPKC-Par6 cannot be activated by Cdc42, but it is hypothesised that subsequent interaction of Par6 with Crumbs could make aPKC competent for Cdc42 activation again (Dong et al. 2019). Together, these interactions promote the accumulation of Par6-aPKC with Crumbs and Cdc42 on the apical surface, while restricting Par3 to adherens junctions.

Moreover, aPKC enforces apical identity by phosphorylating and displacing lateral polarity proteins including Par1, Lgl and and Yurt (Gamblin et al. 2014; Suzuki and Ohno 2006).

It was demonstrated that Yurt can oligomerize through its FA domain and this promotes its binding to FERM domain of Crumbs. Binding of Yurt oligomers inhibits ectopic spreading of Crumbs to the lateral membrane. In turn, aPKC phosphorylation of the FA domain destabilises the oligomer state of Yurt and represses its function (Gamblin et al. 2018). If Yurt also directly or indirectly inhibits aPKC kinase activity is not known.

It was thought that aPKC phosphorylates polarity proteins within the apical domain as well. Phosphorylation of the FERM binding domain of Crumbs was suggested to stabilize Crumbs by inhibiting its endocytosis in experiments including overexpression constructs (Sotillos et al. 2004; Fletcher et al. 2012). However, phosphorylation-resistant Crumbs expressed at endogenous levels is homozygous-viable and has no defects in polarity and development; so even if aPKC does phosphorylate Crumbs in vivo, this phosphorylation is dispensable (Cao et al. 2017).

As in the worm embryo, Par1 phosphorylates and excludes Par3 from lateral membranes to stabilise its position at the adherens junctions (Bayraktar et al. 2006; Benton and St Johnston 2003; McKinley and Harris 2012). Lateral polarity proteins, including the members of Scribble complex, are thought not to only act as substrates but also as inhibitors of apical polarity proteins (Yamanaka et al. 2006), however the molecular mechanisms remain unknown with the exception of Yurt described above.

1.5 A novel biophysical view of cell polarity regulation

Traditional genetic and biochemical methods can only offer a limited model of a complex biological process like cell polarity. Nowadays the biophysical approach, which includes

quantitative imaging and mathematical modelling, is becoming crucial for a comprehensive view of living matter, often revealing new mechanisms about the investigated subject.

Recently, a few biophysical studies emerged that suggested a novel mechanistic view of cell polarity regulation that includes spatial distribution and temporal dynamics. These new insights have been facilitated by developing new tools such as single molecule imaging and tracking. While the majority of these studies were conducted in the context of the worm embryo, they offer a good example of bridging super-resolution imaging methods and cell polarity, which is important for understanding the scientific objectives of this thesis.

Cell polarity regulation can be defined as two distinct processes: polarity establishment and polarity maintenance. The latter being the question of maintaining the border between at least two domains: anterior-posterior or apical-lateral.

Firstly, based on recent biophysical studies, it was proposed that cell polarity maintenance is a dynamic process. This has been shown using fluorescence recovery after photobleaching (FRAP) experiments and single-molecule imaging, which demonstrated that polarity proteins exchange dynamically between the cytoplasm, where they diffuse freely, and the cortex, where diffusion is restricted (Goehring et al. 2010; Goehring et al. 2011b).

In other words, it is thought that this dynamic system exhibits bistability. This means that the system has two stable equilibrium states. Consequently, to achieve this bistable dynamics the opposing domains must be balanced. Moreover, the dissociation rates of one or more polarity proteins must have sigmoidal or ultrasensitive dependence on the concentrations of other (interacting) polarity proteins (Semplice et al. 2012; Lang and Munro 2017). Ultrasensitivity would in this case mean a sharp decay of a polarity protein distribution from one domain upon small increase of concentration of a polarity protein from the opposite domain (Lang and Munro 2017; Sailer et al. 2015).

Oligomerization or clustering has been proposed as one cause of ultrasensitivity. How could this work? Oligomers will associate more strongly with the membrane than single monomers because of increased avidity (Lemmon 2008). Consequently, this can affect the dissociation rate of the oligomerizing protein. That clustering of Par3 could be one of the mechanisms for a bistable system was already suggested by simple computer simulations where dimerization was considered (Dawes and Munro 2011).

Interestingly, single-molecule imaging revealed that in the worm zygote, multiple polarity proteins homo- or hetero-oligomerize, including Par2 (Arata et al. 2016; Rodriguez et al.

2017), Par3 (Li and Bowerman 2010; Sailer et al. 2015; Wang et al. 2017; Rodriguez et al. 2017). Par6-PKC3 (Dickinson et al. 2017; Wang et al. 2017; Rodriguez et al. 2017) and CHIN1 (Kumfer et al. 2010; Sailer et al. 2015; Rodriguez et al. 2017).

During polarity maintenance, CHIN1 undergoes clustering at the posterior cortex (Kumfer et al. 2010; Sailer et al. 2015). CHIN1 cluster growth depends on the local concentration of the Par6-PKC3 complex, where higher concentrations inhibit clustering. Similarly, Par1 could inhibit Par3 oligomerization by phosphorylation of monomers as demonstrated in the fruit fly and vertebrate cells (Benton and St Johnston 2003; Mizuno et al. 2003).

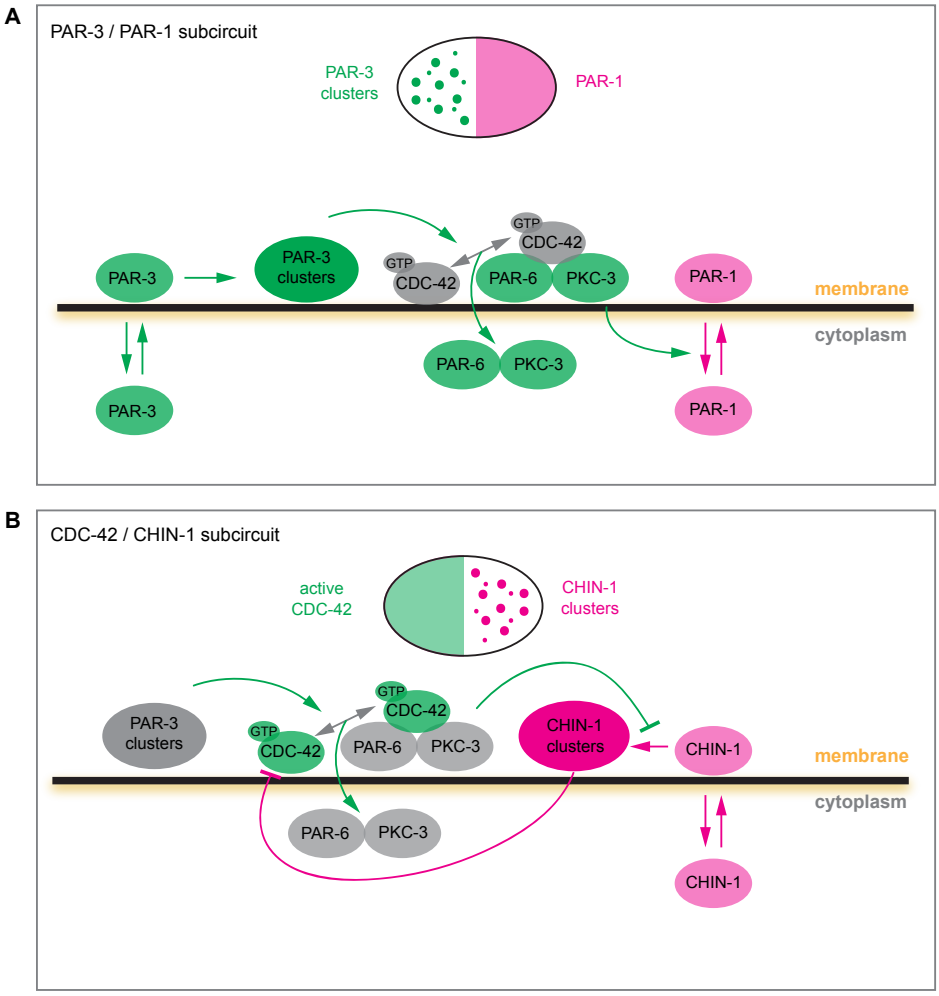


Figure 1.5 **A novel biophysical view of the molecular circuitry between polarity proteins in the worm embryo is based on protein clustering.** Redrawn from (Lang and Munro 2017).

Precise biophysical and genetic experimental tools revealed that anteriorly clustered Par3 recruits active PKC3-PAR6 heterodimers into an inactive cortical complex. From these

clusters, PKC3-PAR6 is recruited then into a complex with active CDC42 and become diffusive and active. Moving posteriorly it inhibits clustering of CHIN1 but upon reaching the critical threshold CHIN1 clusters grow sharply and inhibit CDC42 activity at the posterior. This will affect recruitment of the PKC3-PAR6 complexes.

This clustering therefore acts as a diffusion-based retention system, where cortically immobilised complexes present a sink for freely diffusing cytoplasmic complexes. Consequently, this causes net diffusion from elsewhere and creates asymmetry. In the case of PKC3, there is also a gradient in kinase activity. Interestingly, in the worm embryo both domains contain a polarity protein species that is capable of forming clusters that are ultrasensitive to the concentration increase of a respective kinase from the opposite domain (Figure 1.5). Importantly, deciphering these two molecular circuits would not be possible without combining genetic and biochemical tools with novel single-molecule imaging approaches.

It remains unclear if a similar mechanism to this spatial regulation and molecular circuitry demonstrated in the worm embryo also take place in other systems, like epithelial cells.

1.6 Open questions and scientific objective

Despite extensive work on cell polarity and its regulation in different experimental systems many fundamental questions remain unanswered about epithelial cells. Before I started my PhD, the spatial distribution characteristics of polarity proteins in epithelial cells were poorly understood and not really considered to play an important role in cell polarity regulation. Of course, this was not omitted intentionally since imaging methods were still not powerful and precise enough to image proteins in thick samples and also to appropriately quantify the obtained images. With the development of super-resolution imaging techniques in recent years, the answers to this question became reachable. The objective of the experimental work presented in this thesis was to establish a pipeline to be able to quantitively address the spatial distribution of polarity proteins in epithelial tissue.

To first understand if there are any spatial and temporal patterns of cell polarity proteins in epithelial cells, I assessed the distribution of apical polarity proteins in mono-layered epithelial tissue using standard confocal microscopy in Chapter 3. Preliminary data showed that apical polarity protein exhibit clustered distribution. Therefore I wanted to establish if these clusters of polarity proteins can be imaged in a thick tissue sample using super-resolution microscopy. For this I optimised super-resolution imaging pipeline

using the DNA-PAINT approach in Chapter 3. Furthermore, to validate DNA-PAINT for quantification of the number of protein molecules in tissue samples I used the nuclear pore complex for calibration in Chapter 4. Finally, in Chapter 5 I tried to describe the spatial organisation of polarity proteins in a quantitative manner based on the calibration experiments. Moreover, using computer simulations I validated the accuracy of the quantifications and analysis. Altogether, I provide a super-resolution imaging pipeline for imaging of proteins in thick tissue sample and propose a method to quantitatively analyse the resulting images. Importantly, difficulties, especially with quantifications are thoroughly described.

1.7 Fruit fly follicular epithelium as a tissue model system

In order to achieve the aims stated above, I took advantage of the fruit fly as a model system. The tissue of choice was the follicular epithelium that is found in the egg chambers. The advantages of this model system are briefly summarised below.

The fruit fly is an outstanding model system to study cell biology. To start, the fruit fly is cost-effective to culture and has a fast generation time, making it extremely easy to maintain. Making transgenic lines with endogenously-tagged proteins is relatively easy, since the majority of proteins are encoded by a single gene. Additionally, it is very easy to obtain large amounts of material for experiments. For example, with a quick dissection of a few female flies, one can obtains hundreds of egg chambers. The egg chambers can live ex-vivo for a couple of hours, which allows this tissue to be imaged live and intact.

Each adult fruit fly female contains a pair of ovaries, with each ovary composed of 16 ovarioles. An ovariole is a string of developing egg chambers in different stages. At the anterior tip of the ovariole lies the germarium where the germline stem cells. From the germarium egg chambers bud posteriorly and grow in size until they are ready for fertilisation (Figure 1.6). Developing egg chambers are classified into 14 stages, based on their morphological characteristics (Spradling 1993). At stage 2 they have spherical surface and are around 25 um big, at stage 7 they elongate and have ellipsoid surface with length around 120 um, while at stage 14 they reach 800 um in length and 300 um in width (Jia et al. 2016).

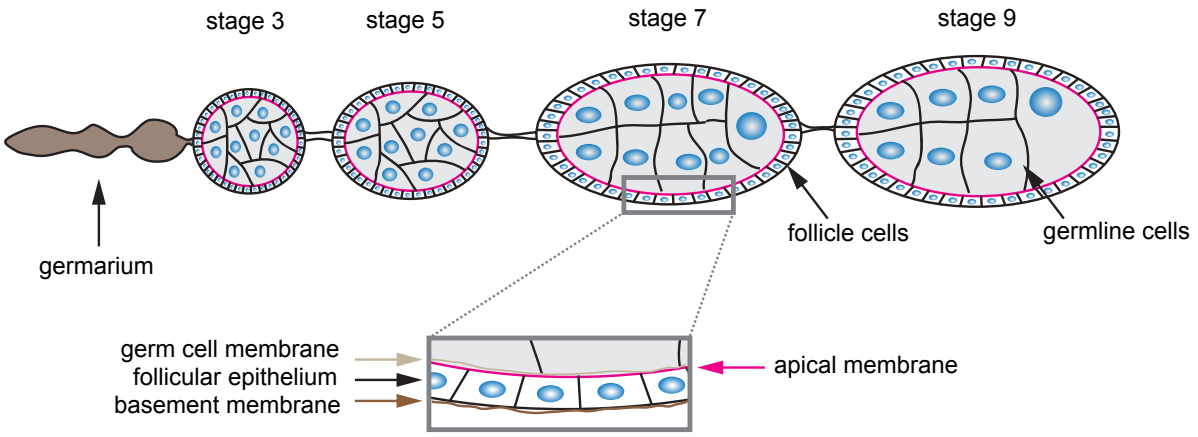


Figure 1.6 **Schematic view of *Drosophila melanogaster* egg chamber.** Each ovariole consist of egg chambers in different developmental stage (stage 1-14). The germ cells are surrounded by a monolayer of follicular epithelium with the apical side facing the germ cell membrane. Redrawn from (Schneider et al. 2006).

Follicle cells form a simple mono-layered epithelium that surrounds each Drosophila egg chamber, and this is one of the most thoroughly investigated epithelia in the context of cell polarity (Tepass 2012). The follicular epithelium is polarised, with apical membranes contacting the oocyte and the basal membrane facing outwards (Figure 1.6). The follicle cells are mitotically active until stage 6 (Deng et al. 2001). They change their shape during egg chamber development. In the early stages, they have a cuboidal shape and their height is around 5-7 um. In the later stages (after stage 9) anterior follicle cells become squamous (cell height around 2-3 um) and posterior follicle cells become columnar (cell height around 7-10um) (Spradling 1993).

A big advantage of this tissue is the possibility of generating mutant clones of cells in a tissue that consist of wild-type cells. This not only allows one to investigate the consequences of otherwise embryonically-lethal mutations, but also to evaluate the background levels when it comes to imaging.

Chapter 2

Materials and methods

2.1 Fly husbandry and stocks

Standard procedures were used for Drosophila husbandry and experiments. Flies were maintained on standard fly food supplemented with live yeast at 25 °C. *yw* flies (Bloomington 1495) were used as a wild-type control unless otherwise specified.

The following Halo- or SNAP-tagged stocks were used: aPKC-Halo, Par6-Halo, Crumbs-Halo, Crumbs-SNAP, Nup160-Halo, Nup188-SNAP, and aPKC-Halo-SNAP. All Halo- or SNAP-tagged fly stocks were generated for this study by Nick Lowe, Jenny Richens and Amandine Palandri (see below). The following fluorescent stocks were used: E-Cadherin-GFP (Huang et al. 2009), Par6-GFP (Wirtz-Peitz et al. 2008), Crumbs-GFP (Huang et al. 2009), aPKC-GFP (Chen et al. 2018), Par3-GFP (Buszczak et al. 2007), UAS-GBP-mKate-CAAX (made by Dmitry Nashchekin, Daniel St Johnston's laboratory).

The following background stocks were used to generate mitotic clones: hs-FLP; FRT82B-GFP was used to generate negatively mark (absence of GFP) Crumbs-Halo clones, hs-FLP; FRTG13-GFP was used to generate negatively mark (absence of GFP) aPKC-Halo clones, FRPnls, hs-FLP, FRT19A was used to generate negatively mark (absence of RFP) Par6-Halo clones. Ectopic expression in the follicle cells was driven by Traffic Jam-Gal4 (follicle cell specific) or Tubulin-Gal4, Gal80(temperature sensitive) (for more sensitive expression).

2.2 Cell culture

U2OS cell lines expressing Nup96-Halo and Nup96-SNAP, respectively, were obtained from Jonas Ries' laboratory (EMBL, Heidelberg) (Thevathasan et al. 2019). The cells were passaged every 4 days. The cells were maintained in DMEM supplemented with 10% fetal bovine serum and 1% penicillin/streptomycin. Passaging was performed using 1x PBS and Trypsin-EDTA 0.05%.

2.3 Genetic techniques

2.3.1 Fly transgenesis using site-specific recombination system

For the ectopic expression of Halo-CAAX flies, UAS/Gal4 system was used (Duffy 2002). The pUASp-attB vector backbone was used to prepare the transgenic construct. The vector was digested using NotI and BamHI restriction enzymes. The Halo-CAAX was then introduced into the backbone vector using Gibson cloning. The vector was injected into 0-1 h old embryos containing attP40 landing site and transformants were selected in subsequent generations on the basis of the red eye colour. The microinjections were performed by John Overton.

2.3.2 Fly transgenesis lines using CRISPR-Cas9 system

Tagging of polarity proteins was designed and performed by Nick Lowe. For endogenous-tagging of the protein targets CRISPR-Cas9 system was used. Unless otherwise noted all reference to left and right arms assumes the direction of transcription is from left to right. Gene sequences were amplified from CFD2 nos-Cas9 fly DNA (Port et al. 2014). Homology arms were TOPO cloned into Invitrogen vectors pCRII TOPO or pCR2.1 with the exception of the Par6 donors where amplified homology arms were cloned as a Bam-XhoI fragment. Guide RNA was produced by in-vitro transcription as described before (Bassett et al. 2013). gRNA and plasmid donor were co-injected into the nanos-Cas9 CFD2 line as described before (Port et al. 2014). Typical concentrations were 80 ng/µl gRNA and 400 ng/µl plasmid donor. Single flies were mated to *yw* flies of the opposite sex, after which the parental fly was rescued and putative productive broods, as determined by PCR of the parental fly, were used to produce subsequent generations. For details about the primers used and the CRIPSR target sequences see Appendix A.

Crumbs-Halo

Halo-tag was inserted in the extracellular domain, close to the transmembrane domain on extracellular surface. 1.5 kb of left and right homology arms flanking the Crumbs transmembrane domain were amplified using the primers *Crb2F* and *Crb2R*. The PCR product was then cloned into pCR2.1 vector. Halo-tag was obtained by PCR amplification of Halo-tag with primers *HaloCrumbsF* and *HaloCrumbsR*, then inserted into previously mentioned pCR2.1 vector that was linearized using primers *CrbtagR* and *CrbtagF*. Insertion of the Halo-Tag is just 5' of the transmembrane domain and disrupts the CRISPR target.

Par6-Halo

Halo-tag was inserted at the C-terminus of the protein. 2 kb left and right homology arms were amplified with primers *Par6Kpn1p* and *Par6Not1m* and cloned into pBluescript SK cut with KpnI and NotI. Silent mutations were introduced into the CRISPR target site using primers *Par6CRF1mutF* and *Par6CRF1mutR*. This plasmid was linearized by PCR using primers *pBS-Par6-vec-rev* and *pBS-Par6-vec-fwd*. The Halo-tag was amplified using primers *HaloTag-fwd* and *HaloTag-rev*, and then Gibson assembled into the plasmid.

aPKC-SNAP

SNAP-tag inserted at the N-terminus of the protein. Contains 1.2kb left and 2kb right homology arms either side of the translation start site. Genome was amplified with primers *aPKC5UT1p* and *aPKCint2m*. Initial plasmid construct was made with a GFP insertion by overlapping PCR with *GFPaPKCfus2m* and *GFPaPKCfus1p* to amplify GFP and mutate the CRIPSR target together with *aPKCCRF1mut* and *GFPaPKCfus1m*. This GFP-aPKC plasmid (N996) was linearized using primers *aPKCnt-rev* and *aPKCnt-fwd*. The SNAP-tag insert was amplified with primers *SNAPaPKC-fwd* and *SNAPaPKC-rev*, and then Gibson assembled into the plasmid.

aPKC-Halo

As with the creation of the SNAPaPKC donor, the plasmid pCRII/GFPaPKC was linearized using primers *aPKCnt-rev* and *aPKCnt-fwd*. HaloTag was amplified with primers *Halo-aPKC-fwd* and *Halo-aPKC-rev*, before Gibson assembly.

Nup160-Halo

Tagging of polarity proteins was designed and performed by Jenny Richens and Amandine Palandri. Halo-tag was inserted at the C-terminus of the protein. Genome was amplified in two portions. The first portion with primers *Nup160Halo-F1* and *Nup160Halo-R1*. The second portion with primers *Nup160Halo-F2* and *Nup160Halo-R2*. Both inserts together with the Halo-tag sequence were inserted into pBluescript SK(+) plasmid cut with EcoRI and NotI. Two pCFD3 plasmids containing two different CRIPSR target sequences were together with a plasmid donor injected into the nanos-Cas9 CFD2 line.

Nup188-SNAP

Tagging of polarity proteins was designed and performed by Jenny Richens and Amandine Palandri. Halo-tag was inserted at the C-terminus of the protein. Genome was amplified with primers *Nup188SNAP-F* and *Nup188SNAP-R*. The genomic insert and the SNAP-tag sequence were inserted into pBluescript SK(+) cut with EcoRI and NotI. Two pCFD3 plasmids containing two different CRIPSR target sequences were together with a plasmid donor injected into the nanos-Cas9 CFD2 line.

2.3.3 Generation of wild-type cell clones in epithelial tissue

Homozygous mitotic clones expressing the Halo- or SNAP-tagged polarity protein within the wild-type tissue were generated using the FLP/FRT mediated system (Xu and Rubin 1993). In brief, a fly stock containing an FRT site upstream of the tagged polarity protein loci was crossed with the respective stock containing an FRT site upstream of the gene for the nuclear GFP. L3 stage larvae and pupae were heat shocked at 37 °C for 2 h, twice a day for 3 days. After the emergence of the adult flies, the vials were transferred to 25 °C for 2 days before the ovaries were dissected and prepared for imaging.

2.3.4 DNA preparation from single flies for PCR

A single fly was placed in a 0.5 ml tube and mashed for 5-10 s with a pipette tip followed by adding 50 µl of squishing buffer. The sample was incubated at 37 °C for 30 min and then heated for 2 min at 95 °C to inactivate the Proteinase K. The extract was used for PCR or stored at −20 °C until needed. The squishing buffer was made as following: 10 mM Tris-HCl (pH 8.2), 1 mM EDTA, 25 mM NaCl, just prior to use Proteinase K was added to a final concentration of 200 µg/ml.

2.4 Biochemical techniques

2.4.1 Materials

Amicon Ultra 0.5 ml centrifugal filter units (cat: UFC503024) were purchased from Sigma Aldrich. SNAP-tag polyclonal antibody (cat: P9310S) was purchased from New England Biolabs. Halo-tag monoclonal antibody (cat: G921A) was purchased from Promega. 20X Bolt MOPS SDS running buffer (cat: B0001) was purchased from Thermo Fisher. Bolt 8% Bis-Tris Plus gels (cat: NW0082BOX) were purchased from Thermo Fisher. Protein deglycosylation kit (cat: P6044S) was purchased from New England Biolabs. Immobilon-FL membranes (cat: IPFL00010) were purchased from Millipore.

2.4.2 Buffers

2.4.3 Nup96 protein extract preparation from cultured cells

Cells were seeded in a 6-well dish and left until they reached the confluency. Then they were prefixed in 2.4% PFA for 30 s, followed by permeabilization in 0.4% Triton X-100 for 3 min. They were then fixed in 2.4% PFA for 30 min, followed by washing twice in PBS for 5 min. Cells were then labelled for 2 h at room temperature on a shaker. The labelling solution consisted from 1 µM of Halo- or SNAP-ligand conjugated to the docking oligo and 1 µmol of DTT in 0.5% BSA (in the case of the non-labelled control, only 0.5% BSA was used). After labelling cells were washed twice in PBS for 5 min. The dish was placed on ice and kept chill for the further steps. Cells were washed twice with ice cold PBS. Then cold customised RIPA buffer (for composition see Table 2.1) with protease inhibitors was added (250 µl per well) and incubated on ice for 5 min, swirling the plate occasionally for uniform spreading. Cells were then scraped with a cell scraper and the lysate was transferred to a microcentrifuge tube. The lysate samples were spun down at 13000 rpm for 15 min to pellet the cell debris. 1 glass bead was put into each tube and the sample was sonicated for 15 min in sonicating water bath at 4 °C. The sample was then boiled for 30 min at 100 °C and then spun down for 10 min at 13000 rpm. The supernatant was then mixed with appropriate amount of 2x Laemmli buffer. The sample was boiled for 10 min at 95 °C prior to loading on gel.

Buffer	**Composition**
1X NuPAGE/Bolt MOPS SDS running buffer (500 ml)	25 ml of 20X running buffer and 475 ml of dH2O
10X transfer buffer (1l)	179.6 g of glycine, 37.5 g of Tris base, 31.2 ml of 10% SDS, make up to 1l with dH2O
1X transfer buffer (1l)	100 ml of 10X transfer buffer, 700 ml of dH2O, 200ml of methanol
1X TBS (100 ml)	10 ml of 10X TBS, 90 ml of dH2O
1X TBS-T (TBS, 0.1% Tween) (1l)	100 ml of 10X TBS, 5 ml of 20% Tween, 895 ml of dH2O
Blocking buffer (1% BSA, 3% skimmed milk powder, TBS, 0.1% Tween) (250 ml):	2.5 g of BSA, 7.5 g of skimmed milk powder, 25 ml of 10X TBS, 1.25 ml of 20% Tween, make up to 250 ml with dH2O
Secondary Antibody Diluent for LI-COR (30 ml)	add Tween-20 to a final concentration of 0.2% and SDS to a final concentration of 0.01%
Wet transfer buffer (1l)	5.8 g of Tris/Trizma, 11.3 g of glycine, 200 ml of methanol and 1 mL of 20% SDS, make up to 1l with dH2O
Customised RIPA buffer	150 mM NaCl, 1% IGEPAL CA-360, 0.5% sodium deoxycholate, 2% SDS, 300 mM Tris-HCl (pH 7.4)

Table 2.1: Buffers used in this work.

2.4.4 Nup160 protein extract preparation from fly ovaries

Fly ovaries (5 pairs per sample) were dissected in Schneider's medium at room temperature and then fixed in 4% PFA in PBS for 20 minutes. After washing in PBS 3 x 5 min, the ovaries were permeabilised in 0.5% Triton X-100 for 10 min, followed by a quick rinse in PBS. The sample was then labelled for 1 h at 37C while shaking at 550rpm. The labelling solution consisted from 1 µM of Halo- or SNAP-ligand conjugated to the docking oligo in 0.5% BSA (in the case of the non-labelled control, only 0.5% BSA was used). After labelling the sample was washed in PBS 4 x 10 min. The liquid was then removed from the tube and 50 uL of modified RIPA buffer was added (10 µl per pair of ovaries). The sample was then incubated for 20 min at 4 °C. Following by putting 1 glass bead per tube and sonication for 15 min in sonicating water bath at 4 °C. The sample was then boiled for 30 min 100 °C and span down at 13000 rpm for 10 min. The supernatant was then transferred

to a new tube. To further remove low-molecular weight proteins the lysate sample was purified using the filter column (pore size = 30 kDa). The unit was first equilibrated with 500 µl of ddH_2O and span down at 13000 rpm for 15 min at 4C. The flow-through was discarded and the sample was loaded onto column, followed by spinning it at 13000 rpm for 10 min at 4 °C. The filter was then inverted into a new tube and elution was done at 1000 rpm for 10 min at 4 °C. The resulting liquid was then mixed with appropriate amount of 2x Laemmli buffer. The sample was boiled for 10 min at 95 °C prior to loading on gel.

2.4.5 Gel electrophoresis

8% Bolt gels were used, expect for Crumbs-SNAP when 3-8% Tris-Acetate gel was used. For Crumbs-SNAP the gel was run for 72 h at 120 V at room temperature or at at 4 °C. For aPKC-Halo the gel was run for 21 h at 85 V at room temperature. For Par6-Halo the gel was run for 6 h at 85 V at room temperature. For Nup160-Halo the gel was run for 6 h at 110 V at room temperature. For Nup96-Halo or Nup96-SNAP the gel was run for 17 h at 55 V at 4 °C or for 8.5 h at 100 V at room temperature.

2.4.6 Quantitative Western blot (gel band shift assay)

For determining absolute labelling efficiency (ALE) using the gel band shift assay the gels were first transferred to Immobilon-FL membranes for 45 min at 16 V, 0.5 A using wet transfer. The membrane was then blocked in blocking buffer for 1 h at room temperature or overnight at 4 °C. The blocked membrane was incubated with primary antibody for 1 h at room temperature or overnight at 4 °C. The membrane was then gently washed twice in TBS-T, followed by washing in TBS-T 2 x 5 min and then 2 x 10 min on a shaker. Then the IRDye-conjugated secondary antibody was diluted in IRDye buffer. The membrane was then incubated with the secondary antibody for 1 h at room temperature on a shaker. After that the membrane was gently washed twice in TBS-T, followed by washing in TBS-T 2 x 5 min and then 2 x 10 min on a shaker. Finally, the membrane was rinsed and left in TBS.

For detection of fluorescence intensity of the bands the LI-COR system was used. For ALE quantification I used the Equation 2.1. The background signal refers to fraction of the smear signal intensity above the band of the unlabelled sample. For compositions of used buffers please see Table 2.1.

$$\text{ALE} = \frac{\text{labelled fraction}}{\text{unlabelled fraction} + \text{labelled fraction} + \text{BG}} \tag{2.1}$$

In the Equation 2.1 BG stands for the background signal intensity. The BG value was subtracted from each labelled fraction band intensity. The BG value was calculated for each gel independently.

2.5 Confocal microscopy

2.5.1 Materials

16% formaldehyde (cat: 28908) was purchased from Thermo Fisher Scientific. Carl Zeiss immersion oil (cat: 10539438) was purchased from Fisher Scientific. Schneider's insect medium (cat: S0146) was purchased from Sigma Aldrich. Chambered coverslips (cat: 80826) was purchased from Ibidi. Vectashield (cat: H-1500) was purchased from Vector Laboratories.

2.5.2 Fixed sample preparation

Fly ovaries were quickly dissected to isolate ovarioles in 4% PFA in PBS at room temperature and then further fixed while rotating for 20 min overall. The sample was then washed 3 x 5 min in PBS. The ovaries were then put in Vectashield mounting medium overnight at 4 °C and mounted next day on a microscope slide.

2.5.3 Live sample preparation

For live imaging fly ovaries were dissected to isolate ovarioles in Schneider's medium at room temperature and the muscle sheet surrounding the egg chambers was removed. The ovarioles were then put in a chambered coverslip.

2.5.4 Optical setup

Image acquisitions were carried out on an inverted Leica SP5 microscope. For all experiments an oil-immersion 63x NA 1.40 CS2 objective was used. For GFP 485 nm laser wavelength was used and for mKate 588 nm laser wavelength was used, both at 9% laser

power. PMT camera was used for detectin with 1250% smart gain and 6X zoom, pixel size was 30 nm.

2.5.5 Image data post-processing

Integrated signal anisotropy (ISA) was calculated in the confocal images of cell junctions. ISA was calculated by multiplying standard deviation of mean pixel value along the junction with the frequency of bright peaks along the junction. Higher ISA value means less homogenous fluorescent signal along the junction. The calculation was performed using custom written Fiji plugin (see Appendix B).

For obtaining the recovery curves in FRAP experiments I performed a double normalisation as previously reported (Gomez-Lamarca et al. 2018). For each experiment the fluorescence intensities in bleached (B), unbleached (UB), and background (BG) regions were measured over time using ImageJ. The bleached intensity value was measured in the junctional area, which fluorescence was bleached by the high-power laser pulse illumination. The unbleached intensity value was measured in the neighbouring junctional area, which fluorescence was not bleached by the high-power laser pulse illumination. The background intensity value was measured in the cytoplasmic area. In a double normalisation approach the signal in the bleached region is normalised to the average prebleach signal and the background signal, as following: $((B - BG)/(B_{pre} - BG)) * ((UB_{pre} - BG)/UB - BG)$.

2.6 Super-resolution microscopy

2.6.1 Reagents

Unmodified and modified (conjugation to ligand or fluorophore) DNA oligonucleotides were purchased from Integrated DNA Technologies or AtdBio. All DNA oligonucleotides (conjugated to ligand or fluorophore) were kept as 100 µM stock solutions in nuclease free water. DNA origami structures were purchased from Gattaquant DNA Nanotechnologies. A non-commercial modified version of Halo ligand (PBI 300-43) was a gift from Mark McDougall at Promega. SNAP ligand (cat:S 9148S) was purchased from New England BioLabs. Phalloidin-iFluor 405 (cat: ab176752) was purchased from Abcam. AF647-BG (SNAP-Surface Alexa Fluor® 647, cat: S9136S) was purchased from New England Biolabs. JF646-CA was a gift from Luke Lavis laboratory at Janelia Research Campus.

Bovine serum albumin (cat: BP9702-100, lot: 171120-0162) was purchased from Fisher Scientific. DTT (cat: D11000) was purchased from Melford. Triton X-100 (cat: T8787) was purchased from Sigma Aldrich. Image-iT FX signal enhancer (cat: I36933) was purchased from Thermo Fisher. The microscope slides (cat: 631-0909) were purchased from VWR. The cover glasses (cat: 0107052) were purchased from Marienfeld Superior. Glass-bottomed dishes for seeding and imaging cells (cat: 81158) were purchased from ibidi. Nuclease free water (cat: B1500L) was purchased from New England BioLabs. Catalase from bovine liver (cat: C40-100MG) was purchased from Sigma Aldrich. Glucose oxidase from *Aspergillus niger* (cat: G2133-50KU) was purchased from Sigma Aldrich. 2-mercaptoethanol (cat: M3148-25ML) was purchased from Sigma Aldrich. NH_4Cl (cat: A9434-500G) was purchased from Sigma Aldrich. DMF (cat: D4551) was purchased from Sigma Aldrich. Dithiothreitol (cat: D9779-250MG) was purchased from Sigma Aldrich. Abbelight dSTORM buffer was purchased Abbelight.

2.6.2 DNA-PAINT docking and imager sequences

The DNA sequences used for the docking and the imager oligos were described previously (Schnitzbauer et al. 2017).

Description	**Sequence**
P1 docking oligo	5'-TTATACATCTA-3'
P1 imager oligo	5'-CTAGATGTAT-fluorophore
P3 docking oligo	5'-TTTCTTCATTA-3'
P3 imager oligo	5'-GTAATGAAGA-fluorophore

Table 2.2: DNA-PAINT docking and imager sequences.

2.6.3 DNA origami experiments

DNA origami structures were designed and produced by Gattaquant DNA Nanotechnologies. The structures had 6 single binding sites with P1 sequences (Table 2.2) that were 40 nm apart from each other. The structures were immobilised on a coverslip using a custom-made flow chamber as previously described (Schnitzbauer et al. 2017). The immobilisation was done by first washing the flow chamber 3 x 500 µl with PBS. Then the flow chamber was filled with 200 µl of BSA-biotin solution (1 mg/ml in PBS) for 5 min. The BSA-biotin

solution was then removed by washing it 3 x 500 µl with PBS. The flow chamber was then filled with 200 µl of neutravidin solution (1 mg/ml in PBS) for 5 min. The neutravidin solution was then removed and the chamber was washed 3 x 500 µl with PBS supplemented with 10 mM of magnesium chloride (immobilisation buffer). The DNA origami solution was then diluted with 200 µl of immobilisation buffer. The flow chamber was then incubated with this solution for 5 min. Finally, the flow chamber was washed 3 x 500 µl with immobilisation buffer, sealed and immediately imaged. The optical setup used was the same as described in Section 2.6.9.

2.6.4 DNA oligonucleotide conjugation to ligand

The conjugation of DNA oligonucleotide to Halo or SNAP ligand, respectively, was performed by AtdBio (School of Chemistry, University of Southampton). For Halo ligand I discovered that the commercial version sold by Promega loses its reactivity upon conjugation to DNA oligonucleotide. Mark McDougall at Promega kindly sent us modified version of the ligand (PBI-300-43) (Figure 2.1) fused to C12 spacer. The conjugation of this modified ligand to 5' end of DNA oligonucleotide was performed as following. The reaction solution containing 130 µl of 1 mM of DNA oligonucleotide, 15 µl of 1 M $NaHCO_3$, 10ul of DMF and 10 µl of 100 mM of modified ligand was incubated at room temperature for 2 days. The conjugated DNA oligonucleotide was then purified with reverse phase HPLC, fractions were collected and desalted.

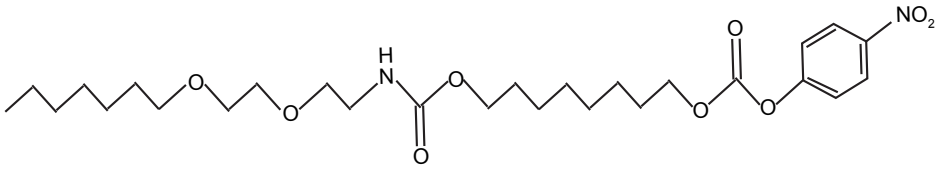


Figure 2.1 **Chloroalkane PBI-300-43.**

2.6.5 Protein labelling in U2OS cells

In U2OS cells proteins were labelled as previously described with a slightly adjusted protocol (Thevathasan et al. 2019). Cells were seeded in a 35 mm imaging dish with a glass bottom 24 h before the labelling, so that they reached 50% confluency next day. First they were prefixed for 30 s in 2.4% (w/v) PFA in PBS and then permeabilised for 3

min in 0.4% (v/v) Triton X-100 in PBS. The main fixation followed for 30 min in 2.4% (w/v) PFA in PBS before incubating for 5 min in 100 mM NH_4Cl in PBS. Cells were then washed 2 x 5 min in PBS and incubated for 30 min in Image-iT FX signal enhancer. After this the cells were labelled with respective ligand for 2 h at room temperature. The labelling solution contained 1 µM of ligand and 1 µM of DTT in 0.5% BSA in PBS. After the labelling the cells were washed for 3 x 5 min in PBS.

2.6.6 Protein labelling in fruit fly egg chambers

Fly ovaries were dissected in Schneider's medium at room temperature and the muscle sheet surrounding the egg chambers was removed. The sample was then fixed in 3% PFA in 0.5X PBS for 15 min, following by 3 x 5 minutes washes in PBS. The egg chambers were then permeabilised in 0.5% Triton X-100 for 10 min. The sample was then incubated in 50 mM ammonium chloride for 5 minutes to quench free aldehyde groups that cause additional fluorescent background when imaging. After a wash in PBS for 5 minutes the sample was incubated in Image-iT FX signal enhancer for 30 min. Immediately after that protein labelling was performed by incubating the sample in 1 µM of either Halo- or SNAP-ligand conjugated to a docking oligo in 0.3% BSA in PBS with 1 µM DTT for 1 hour at 37 °C. The sample was then washed 3 x 5 min in PBS and then washed in 0.1% Triton X-100 overnight. In case of the imaging of polarity proteins, Phalloidin-iFluor 405 (1:500) was added to label the cell cortex (F-actin). Next day the sample was washed 3 x 5 minutes in PBS and then 2 x 5 min in the imaging buffer (500 mM NaCl, pH 7.2), before being mounted on the slide.

2.6.7 Sample mounting

Firstly, the objective slides and the coverslips where cleaned using ether and dried out in a dust-free box. The egg chambers were transferred from the imaging buffer to the freshly prepared imager solution that contained the imager oligo with the enzymatic oxygen scavenging system in the imaging buffer. The sample was then transferred in a 15 µl drop to the objective slide where egg chambers older than the stage 7 were removed. Using the dust-free air all dust particles were blown away from the cover glass (22 x 22 mm) before putting it onto the sample. The cover glass was sealed using a two-compound silicone gel.

The enzymatic oxygen-scavenging system was prepared as following. Base buffer: 1.95 ml of Nanopure water, 2.5 ml of 20% glucose, 250 µl of 1 M NaCl and 250 µl of 1 M Tris

pH 8.0. Base buffer was stored at −20 °C. For imaging the working buffer was freshly prepared by mixing 978 µl of base buffer, 10 µl of catalase (500 kU/ml), 2 µl of glucose oxidase (13.5 kU/ml) and 10 µl of 2-mercaptoethanol.

2.6.8 Optical set up and imaging conditions for U2OS cells

Imaging of the nuclear pore complexes in U2OS cells was carried out on a Nikon STORM (N-STORM) system with an Agilent laser bed. For all experiments an oil-immersion CPI Plan Apo 100X 1.49 NA objective was used. Fluorescence light was spectrally filtered with a bandpass emission filter and imaged pm an iXon Ultra 897 EMCCD camera (Andor). The setup was controlled by Nikon NIS-Elements AR software version 4.50 with N-STORM module.

Unless otherwise stated all acquisitions were acquired with 200 ms exposure time. The size of field of view was 256 x 256 pixels, equivalent to 39.9 x 39.9 µm. The pixel size was 160 nm. Cy3B fluorophore was excited with 561 nm laser light wavelength.

2.6.9 Optical setup and imaging conditions for fruit fly egg chambers

Image acquisitions were carried out on an inverted Olympus microscope. For all experiments a silicon oil-immersion objective (Olympus, 100X UPlanSApo 1.35NA) was used. The microscope room was located in the basement of the Gurdon Institute. The average ambient temperature was 22 °C.

Fluorescence light was spectrally separated and filtered with appropriate dichroic and emission filters. Specifically, two filter set were used as follows. First, the microscope has a filter set composed of a quad-band dichroic mirror (Chroma, ZT405/488/561/647rpc) and emission filter (Chroma, ZET 405/488/561/647m) in the turret for imaging with blue, green, orange and far red fluorophores. Second, the system has a two-channel filter set more amenable to DNA-PAINT imaging composed of a two-band dichroic (Chroma, 59007bs) and corresponding emission filter (Chroma, 59007m) also in the microscope turret for imaging Cy3/Cy5, or equivalent fluorophores. After leaving the microscope base, fluorescence emission is further split with a custom dual channel imaging splitter sending far red light to one portion of a camera sensor and the remaining light to an adjacent area of a camera sensor (Chroma, ZT647rdc). The latter light path includes a motorized filter wheel such that the

desired colour channel may be selected (Chroma, ET525/50m, ET590/50m, ET705/72m). Images were recorded on a scientific complementary metal–oxide–semiconductor (sCMOS) camera (Hamamatsu, Orca Flash 4.0 V2).

All acquisitions were acquired with 250 ms exposure time, however as the camera was operated in "light sheet" mode, introducing an effective confocal slit to reduce out of focus background, the per pixel exposure time was significantly reduced to approximately 10 ms. The effective camera pixel size was 97 nm. A 256 x 512 ROI was set on the camera and two colour channels were projected, respectively, onto the left and right side of this ROI. Each colour channel had an image format of 256 x 256 pixels (half the ROI width). Only a 200 x 200 pixel area of the respective colour channels was used and resulted in a field of view of 20 x 20 µm.

Cy3B was excited with 561 nm laser light and imaged with the aforementioned filter set. Unless otherwise stated, 561 nm laser excitation would be 0.6 kW/cm^2 for epi fluorescence illumination. Atto655 was excited with 642 nm laser light wavelength using the aforementioned quadband emission filter for single-color imaging and with the aforementioned Cy3/Cy5 emission filter when imaging Cy3B fluorophore simultaneously. Unless otherwise stated, 642 nm laser excitation would be 1.2 kW/cm^2 for epi fluorescence illumination. Please not that above stated measurements assume a uniformly illuminated area, while in line scanning setup the average energy per unit area per unit time will differ.

To localise the marginal zone in the follicle cells, phalloidin staining that labels the cell cortex was used.

2.6.10 Drift correction

As egg chambers are not adhered to anything (that is, they are floating in solution) drift correction is critical for DNA-PAINT imaging where acquisition times routinely range from tens of minutes to hours. While sample XY drift can be address in data post-processing, sample drift in z-axis must be corrected in real time. Here, a real time drift tracking and correction algorithm based on transmitted light images was implemented based on the previous work (Mcgorty et al. 2013). In this scheme, references images are collected above, at, and below the focal plane prior to image acquisition. Once image acquisition has start, transmitted light images in a near-infrared (NIR) colour channel are captured on a camera separate from the fluorescence detection camera. These real time NIR images are

compared with the pre-acquired reference images and from normalized cross correlation the sample position can be determined in XYZ relative to the initial position. Here, sample z drift was corrected in real-time every 0.8 s.

Sample XY drift was corrected in the post-processing using Matlab script with redundant cross-correlation algorithm (Wang et al. 2014), using first 1000 frames for the cross-correlation with the next 1000 frames, etc.

2.6.11 Image data post-processing

Acquired frames from the Olympus setup were processed in Matlab using custom-written script by Yongdeng Zhang (Joerg Bewersdorf laboratory, Yale University). Only the blinks above 500 photos were fitted and localised, maximum threshold was 8. Single-molecule candidates from the merged frames were isolated and fitted with an elliptical Gaussian model using a maximum likelihood estimator accounting for the camera-specific noise associated with sCMOS cameras (Huang et al. 2013).

Acquired frames from the Nikon setup were processed in Picasso software (Jungmann et al. 2016). The parameters used for fitting were as following. For blink identification the box side length was set to 7 and the minimum net gradient was set to 64000. For photon conversion the EM gain was set to 300, baseline was set to 100, sensitivity was set to 1, quantum efficiency was set to 0.80. The binding kinetics was processed in Picasso software as well.

2.6.12 Image data analysis

For determining the percentage of localisations per 1 µm^2 due to non-specific binding events I analysed super-res images of aPKC-Halo in follicular tissue that also contained clones of wild-type cells. I first quantified the following values in the tagged-cell area: the number of cytoplasmic localisations (CL), the number of junctional localisations (JL), the cytoplasmic area in 1 µm^2 (CA), the junctional area in 1 µm^2 (JA), and the ratio between JA and CA, also known as the area ratio (AR). In the wild-type clone area I quantified the number of localisations per 1 µm^2, also known as background localisations (BL). I then used the following Equation 2.2:

$$\%\text{ of background localisations per } 1\,\mu m^2 = \frac{\text{BL x (CA + JA)}}{\text{(JL x AR + ((CL x (1 - AR))}} \tag{2.2}$$

The statistics of the imager oligo binding kinetics was analysed using Picasso software that was developed by Ralf Jungmann laboratory (Jungmann et al. 2016), The Matlab file with the coordinates of localisations points and their respective frame numbers was transformed into hdf5 file type that was used in Picasso software.

For counting the molecules custom-written Matlab script was used. Region of interests (ROIs), e.g. junctions, cytoplasmic areas, NPCs, were defined in Picasso and exported as yaml files. Junctional ROIs were always 250 nm wide and the length depended on the junction. Cytoplasmic ROI was $1\,\mu m^2$ big circle. NPC ROI was $1256\,nm^2$ big circle for the subunit and $45\,238\,nm^2$ big circle when the entire NPC was analysed.

For predefined ROIs either: only the number of molecules per ROI was calculated or clustering analysis was additionally performed (for details see Appendix H and Appendix I). In either case, the number of molecules per ROI or per cluster was calculated using the influx rate as determined in Chapter 4. Before the counting of the binding sites the junctional and the cytoplasmic ROIs were cleaned by removing all single localisations and all clusters having less than 10 localisations. The rationale behind this is that upon 20000 frames long acquisition and influx rate of 0.0005387 per frame, a single binding site would produce a cluster of on average 10-12 localisations (depends on the duration of the binding event). The mean t_{OFF} of the ROI or a cluster was determined by calculating the mean of the exponential function that was fitted to cumulative distribution function of the all t_{OFF} for each ROI. The calculated mean t_{OFF} per cluster was used to calculate the number of binding sites per cluster as previously described (Jungmann et al. 2016). The theory behind the calculation is also described in Chapter 4.

For determining effective labelling efficiency (ELE) the super-resolution images of NPCs were automatically segmented using custom written Fiji plugin (see Appendix C). I only used the numbers of NPCs with 4-8 labelled subunits to plot the distribution that was then fitted using custom written Matlab script (see Appendix D).

2.6.13 Statistical analysis

Details of statistical analysis for each experiment are provided in a respective figure legend or in the main text. Analyses were performed in Matlab.

Chapter 3

Establishing the super-resolution imaging pipeline for the fruit fly tissue

3.1 Introduction

Cell biologists have always been seeking to understand the principles of nature in the greatest detail, that is understanding the localisation and behaviour of sub-cellular structures. With the development of light and fluorescence microscopy, this became commonly available. Currently, the most commonly used fluorescence microscopy technique is confocal microscopy. This technique allows one to resolve sub-cellular structures with down to 200 nm spatial resolution.

In order to better understand how the function of polarity proteins relates to their spatial distribution, it is important to understand their mesoscopic organisation. This means how they organise on the level between nano- and micro-scale. Using confocal microscopy, the distribution of polarity proteins has been mainly descriptive and limited to their cellular location along the apical-basal axis. However, how these proteins are organised within the domain that they localise is still not known.

In this chapter I first describe the use of confocal microscopy experiments to obtain the initial observations about distribution of apical polarity proteins in the fruit fly follicular epithelium. I then continue with establishing the super-resolution imaging pipeline to address questions about the mesoscopic organisation of polarity proteins.

3.1.1 Confocal fluorescence microscopy

In conventional fluorescence microscopy the entire sample is illuminated at the same time and the emitted light is collected. This setup is usually referred to as a widefield microscopy. Although the highest intensity is at the focal point of the objective lens of the microscope, there is illumination of other parts of the sample, resulting in background "noise," which compromises the quality of the image (Figure 3.1A). The quality is usually reflected by resolution that the microscope system can achieve. The resolution is defined as the shortest distance between the two points that can still be distinguished as separate objects.

During the detection process, emitted light rays from a point on the object plane converge to a single point at the image plane. However, the diffraction of light, causes a point on the object plane to blur into a finite-sized spot in the image plane. The three-dimensional (3D) intensity distribution of a point object imaged with a microscope is called the point spread function (PSF) (Figure 3.1B).

Theoretically, a diffraction-limited microscope with numerical aperture (NA) and light with wavelength λ reaches a lateral resolution of $d = \lambda/(2NA)$. NA is the numerical aperture of the objective defined as $NA = n \sin \alpha$, with n being the refractive index of the medium and α being the half-angle of the light that can enter the objective. This was first described by Ernst Abbe and is now known as the Abbe diffraction limit (Abbe 1873).

Experimentally, the resolution can be determined by measuring the full width at half-maximum (FWHM) of the PSF. Two objects closer than the FWHM will not be able to be resolved because the images of their PSFs will overlap. The width of the PSF is about 2–3 times as large along the z-axis as the lateral width for ordinary high NA objectives. A commonly used oil immersion objective with NA = 1.40 results in PSF with a lateral size of $\approx$ 170 nm and an axial size of $\approx$ 425 nm in a refractive index-matched medium and using light with $\lambda \approx$ 480 nm. However, this is only a theoretical PSF where background fluorescence that decreases the PSF intensity is not considered.

The first step towards increased spatial resolution of fluorescence microscopy was made in 1957 when Marvin Minsky patented the confocal imaging technique. Here point illumination instead of widefield illumination is used and a pinhole to eliminates out of focus signal (Pawley 2006). This makes the PSF more narrower and hence increases the spatial resolution (Figure 3.1).

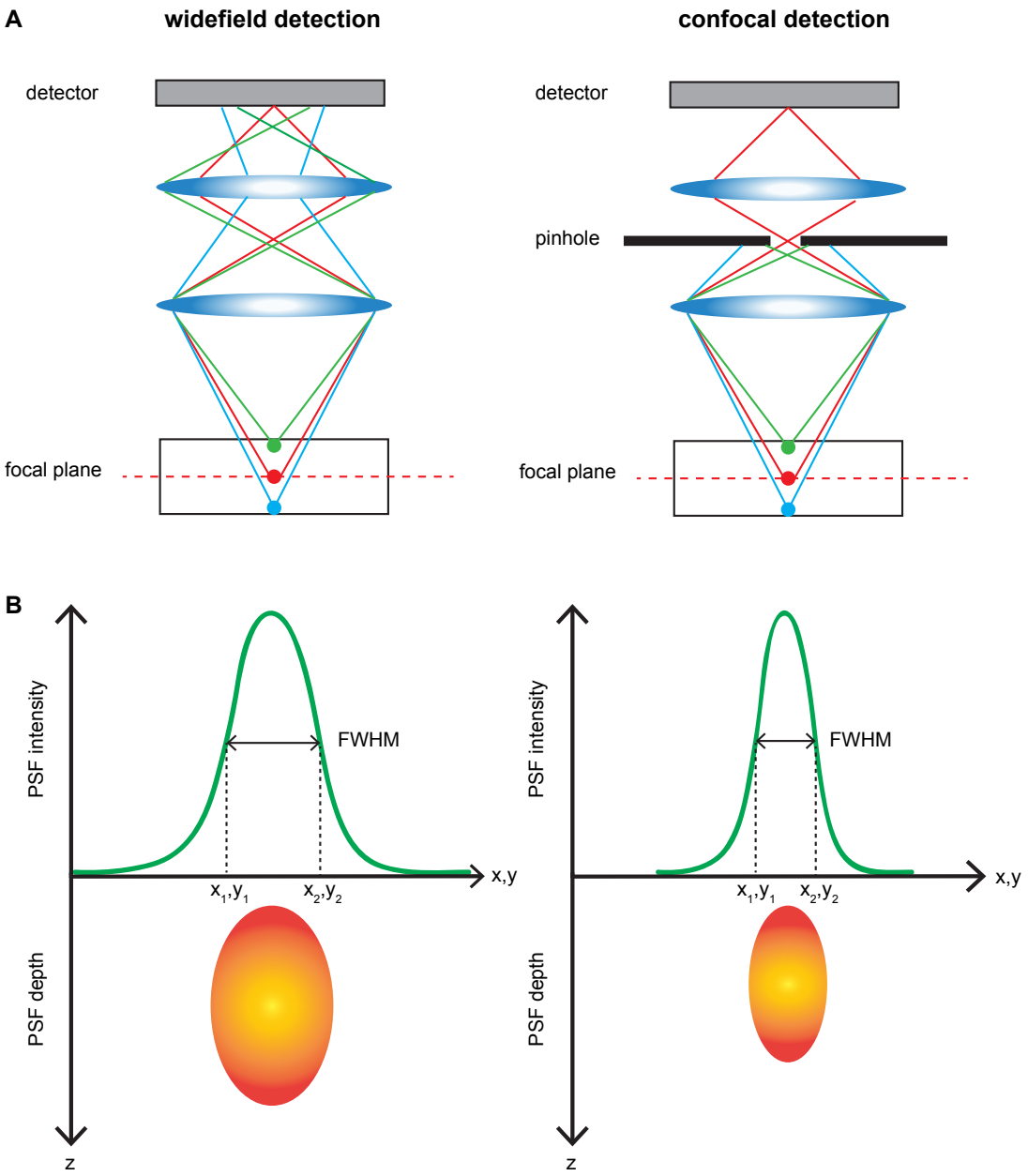


Figure 3.1 **Relation of point spread function (PSF) to widefield and confocal detection.** (A) Left: schematic representation of widefield detection. Light originating from above and below the focal plane (blue circle) will be also detected. Right: schematic representation of confocal detection. Light originating from above and below the focal plane (blue circle) will be blocked by the pinhole, whereas light (red) from the focal plane will be allowed to pass to the detector. (B) Left: axial cross section of the PSF as detected in a widefield setup. The resolution is determine by half-width at full maximum. Right: axial cross section of the PSF as detected in a confocal setup. The resolution is determine by full-width at half maximum (FWHM).

The diffraction limit has been recently pushed to its limit in the context of the confocal microscopy by introducing the Airyscan setup in which extremely small pinholes are

spatialy organized in a particular way to allow for shifting and adding the small signal of each pinhole to a stronger and usable overall signal. In this approach fast-multipixel detectors to reduce the signal loss of small pinhole sizes are used (Huff 2015). Thus the signal-to-noise ratio improves and provides more information for an improvement in resolution by a factor of 1.7 in all spatial directions. This means that you can achieve a resolution of $\approx$ 140 nm laterally and $\approx$ 400 nm axially for $\lambda \approx 550$ nm.

3.1.2 Single-molecule localisation microscopy (SMLM)

I explained that an image of a single fluorophore molecule will result in a diffraction-limited spot, because of diffraction and microscope aberrations. Even if the image were free of effects from lens aberrations, the image would in the best case be the size of a camera pixel (usually 100 nm). However, the image usually spreads over multiple pixels and this image can be fitted with the PSF using a Gaussian function. Nevertheless, the precision of determining the fluorophore position from its image can be much higher than the diffraction limit. This is possible if an image results from multiple photons emitted from the fluorophore (Thompson et al. 2002). Fitting an image consisting of N photons can be viewed as N measurements of the fluorophore position, each with an uncertainty determined by the PSF, thus leading to a localization precision approximated by

$$\Delta_{loc} \approx \frac{\Delta}{\sqrt{N}} \tag{3.1}$$

where Δ_{loc} is the localization precision and Δ is the size of the PSF. The more photons is emitted from the fluorophore, the higher the localisation precision is. This scaling of the localization precision with the photon number allows super-resolution microscopy to circumvent the problem where a resolution is limited by the diffraction of light.

However, when multiple molecules are present in close proximity, localization becomes inaccurate or impossible because the PSF images of these fluorophores overlap. Separation of the fluorescence signal from molecules with overlapping images may be achieved by separating their signal in the time dimension (Figure 3.2A).

Single-molecule localisation microscopy (SMLM) is a collective term that describes different imaging approaches in which single-molecules stochastically switch between bright and dark

states. The differences in SMLM imaging approaches are based on how these bright/dark states are achieved. The three most common approaches are photoactivation localisation microscopy (PALM), direct stochastic optical reconstruction microscopy (dSTORM), and the recently developed DNA point accumulation for imaging in nanoscale topology (DNA-PAINT).

Single molecule localisation microscopy started with the introduction of PALM (Betzig et al. 2006; Hess et al. 2006). In this technique, photoactivatable fluorescent proteins are utilised. At the beginning of the imaging, they are in a non-fluorescent (dark) state. Upon illumination with laser light of 405 nm wavelength a small number of the protein molecules are converted to a fluorescent (bright) state. A laser light with 561 nm wavelength is then used to image the fluorescent proteins and then bleach them back to the dark state. Another cycle of photoactivation follows to image a new subset of proteins until all the proteins are imaged (Figure 3.2B).

dSTORM emerged from the original STORM method. The latter was described upon discovering that a bright (fluorescent) state of a cyanine dye, Cy5, can be pushed into a dark state and this process is reversable. This can be achieved with a red laser light that also induces fluorescent emission from Cy5. The green laser light can be used then to convert Cy5 from the dark back to the bright state, however Cy5 has to be in the proximity of a secondary dye, Cy3. This switching can be cycled between bright and dark states hundreds of times before Cy5 is permanently bleached (Bates et al. 2005; Rust et al. 2006). Soon after the initial description of the STORM technique, a variation called direct STORM (dSTORM) was introduced. This technique uses conventional fluorescent dyes that are able to cycle between fluorescent and dark states upon illumination with laser light of different wavelengths (Heilemann et al. 2008). Importantly, dSTORM does not require an activator dye (Figure 3.2B).

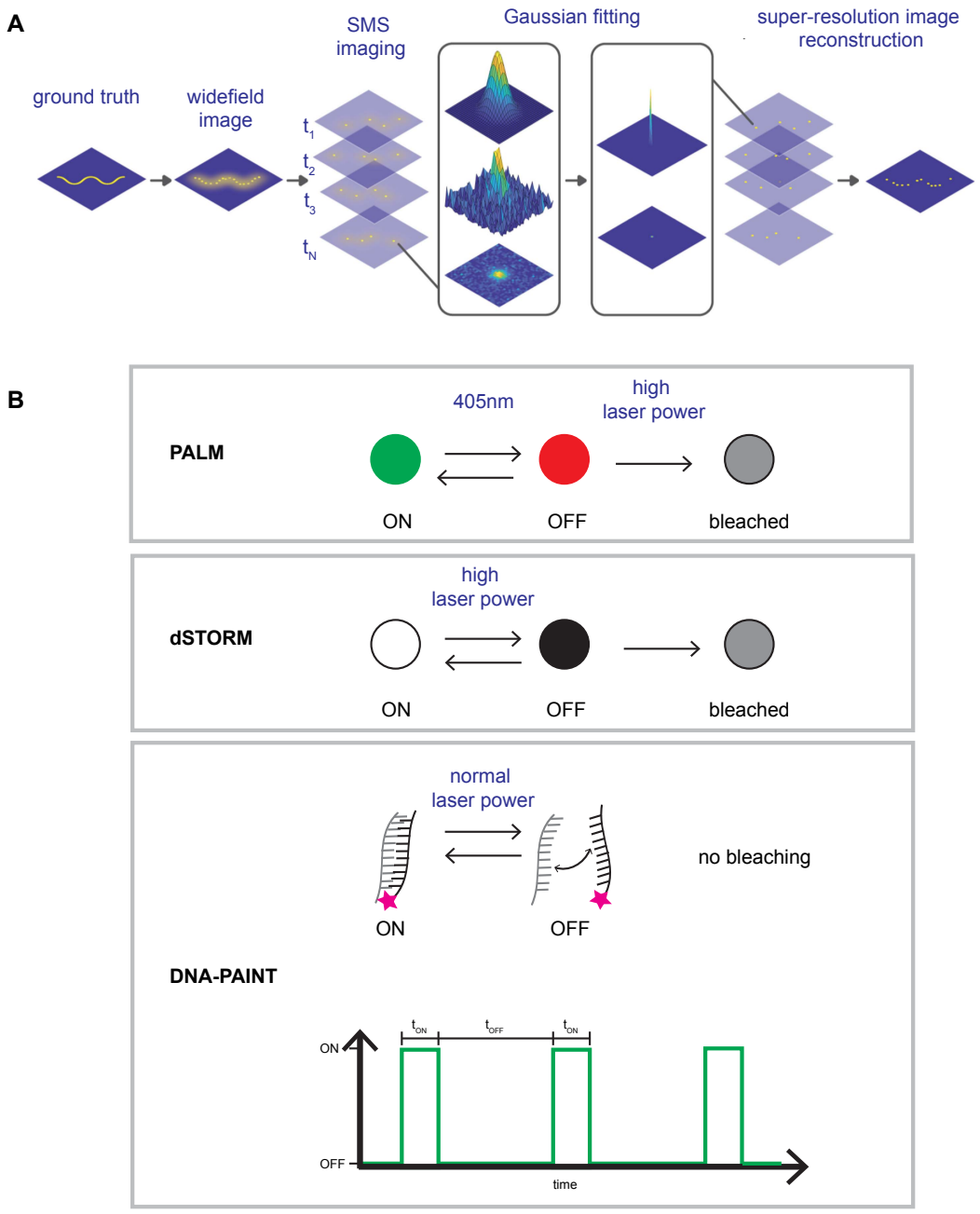


Figure 3.2 **Principles of single-molecule localisation microscopy (SMLM).** Legend on next page.

Figure 3.2 *(previous page)* (A) In wide field imaging resolution is limited because all molecules emit light simultaneously and their PSFs overlap. In SMLM, a subset of single fluorophores is turned on each frame, and this is repeated for thousands of time points. Afterwards each frame is analysed and the centres of each PSF is determined by fitting it with a Gaussian function. By summing al single molecule localizations, that now have a higher localization precision, a super-resolution image is created. Adapted from (Vangindertael et al. 2018). (B) Three different approaches of temporally separating the fluorescent signal in SMLM. Top: In photoactivatable localisation microscopy (PALM) a small subset of molecules is activated by illumination with 405 nm laser light and then permanently bleached with high laser power. Middle: In direct stochastic optical reconstruction microscopy (dSTORM) high laser power pushes fluorescent molecules into a dark state, then they undergo stochastic cycling between the fluorescent and dark state before they get bleached. Bottom: In DNA point accumulation for imaging in nanoscale topography (DNA-PAINT), dye-conjugated oligo transiently hybridises with complimentary docking oligo. Upon hybridization the dye molecule is temporally immobilized, which is detected by the camera. Freely diffusing imager oligos are not detected.

DNA-PAINT

DNA-PAINT was developed in 2010 by Ralf Jungmann (Jungmann et al. 2010). Instead of utilising the photo-switching of the dye, DNA-PAINT achieves the apparent blinking state by the short-time of hybridisation of two single-stranded oligonucleotides (Figure 3.2B). One oligonucleotide is on the target protein and is called the "docking" oligo, while the free-floating fluorescently labelled oligonucleotide is called the "imager" oligo. The free-floating imager oligo cannot be detected by the camera because it is diffusing too fast. Thus the imager oligo can be detected only upon the hybridisation with the docking oligo. The 11-nucleotide long oligonucleotides usually have 9 nucleotides of homology. The time when the imager strand hybridises with the docking strand is called the “ON-time” (t_{ON}), while the “OFF-time” (t_{OFF}) is defined by the length of time between two hybridization events. These two parameters can be adjusted. The "ON-time" can be increased by increasing the strand complementarity, while "OFF-time" can be decreased by increasing the imager strand concentration. Because imager strands are in excess there is no significant photobleaching (Figure 3.2B).

3.2 Experimental design

3.2.1 Confocal imaging of polarity proteins

In order to preliminarily explore the distribution of apical polarity proteins in the fruit fly follicular epithelium, I first utilised confocal fluorescence microscopy for the initial experiments.

The classical view of the polarised epithelial cell has been always pictured laterally (side view), especially in research papers where a fruit fly is used as a model system (Tepass 2012). Imaging cells from the lateral side enables one to observe all three domains at the same time. This view is important when investigating the absence and delocalisation of polarity proteins and their overall effect on the cell and tissue integrity. However, the lateral view is less informative with regard to the spatial distribution of a specific polarity protein within its domain. This is because the x and y axes are actually present a maximum intensity projection of the fluorescent signal along the z axis (Figure 3.3A). Therefore I used an en face (top view) view when imaging apical polarity proteins. They are concentrated in a marginal zone that is less than a micron wide and since an optical section of a confocal microscope covers approximately the same depth, one could argue that with this view all three dimensions are imaged. To localise the marginal zone in the follicle cells, phalloidin staining that labels the cell cortex was used.

Confocal imaging of the fixed endogenously-labelled aPKC, Par6, Crumbs and Par3 revealed that they are not homogeneously distributed along the cell junctions. Rather they formed diffraction-limited areas (Figure 3.3B). In order to check if this is not a fixation artefact, I performed live imaging of the investigated proteins. Also in this case, diffraction-limited areas were observed (Figure 3.4A). This suggests that the non-isotropic distribution pattern of the investigated proteins is of biological origin and not the result of the fixation protocol, which is often the case (Whelan and Bell 2015).

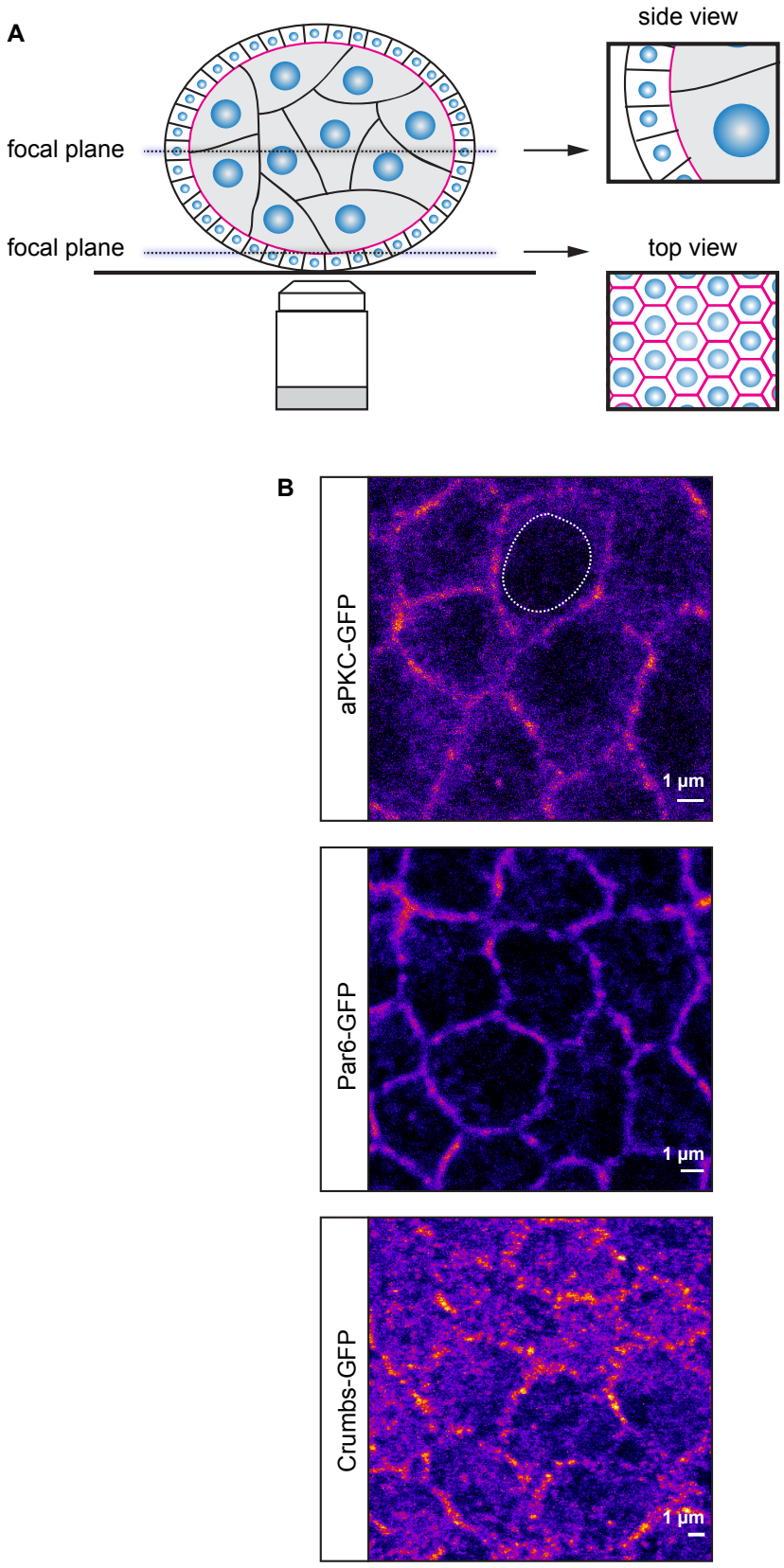


Figure 3.3 **Apical polarity proteins in follicle epithelial cells as seen with confocal microscopy.** Legend on next page.

Figure 3.3 *(previous page)* (A) A schematic of imaging approach of the fruit fly egg chambers. (B) Imaging marginal zone of the follicle cells expressing aPKC, Par6 and Crumbs that were endogenously tagged with GFP. The dashed circle denotes the nucleus position.

To evaluate this fluorescence signal pattern in a quantitative way an integrated signal anisotropy (ISA) was determined using automated image analysis (see Appendix B). ISA can be described as the intensity and frequency of fluorescent signal peaks along the junction. As a negative control mKate-CAAX was used, where the CAAX motif targets mKate to the membrane and the signal appears homogeneous. As a positive control Ecadherin-GFP was used since it is known to cluster and form similar diffraction-limited foci (Truong Quang et al. 2013) (Figure 3.4A). The average levels of the ISA values for the polarity proteins were all statistically significantly larger than for mKate-CAAX, which appeared homogeneously distributed (Figure 3.4B).

Live imaging of endogenously-tagged polarity proteins also revealed that that they are localised in the cytoplasm as well. This was even more obvious when cytoplasmic fluorescence signal is juxtaposed to the nucleus, which lacks GFP signal (Figure 3.3B).

An important component in understanding a protein's spatial distribution is its dynamics. The dynamics of polarity proteins was well explored in the worm embryo where it was demonstrated that they dynamically exchange between the membrane and the cytoplasm (Goehring et al. 2011b). However, their dynamics in the follicular epithelium are unclear, although it is known that they are able to dynamically delocalise during cell mitosis (Carvalho et al. 2015). Previous observations in the St Johnston laboratory using fluorescence recovery after photobleaching (FRAP) methodology showed that aPKC and Par6 recover with a half-life of about 30 s, while Crumbs is relatively immobile (Avik Muhakarjee, PhD thesis). To explore if polarity protein dynamics differs between clustered and non-clustered protein areas and non-clustered protein I performed FRAP on aPKC-GFP.

I bleached a small region in the junctions exhibiting homogeneous fluorescent signal (non-clustered signal) or in the junctions containing bright fluorescent spots (clustered signal) (Figure 3.4C and Figure 3.4D). Within 42 s, the non-clustered fluorescent signal recovered to on average 60% of the initial signal, while this recovery was on average 40% for clustered regions. When the protein being photobleached exhibits 100% mobility, the recovery level should be complete as well. Incomplete recovery of fluorescent signal might

be due to an immobile fraction of the fluorescent protein, whose bleached fluorescent cannot be replenished within the investigated timeframe. To further investigate if the observed recovery rates are really due to immobile protein fractions, present I performed two-step FRAP, where after a first bleaching step the same region is bleached again. If the immobile fraction is real, the second bleach should now only bleach the recovered mobile fraction so the recovery should be complete (Figure 3.4E). I first tested this approach on GFP binding protein (GBP) tagged with mKate and the CAAX motif to target it to the membrane. Its signal appears completely homogeneous along the junctions (Figure 3.4A). After the first bleaching, the signal recovered to an average of 80% of the initial signal, and the same recovery was observed after the second bleach. This incomplete recovery indicates inaccurate normalisation (Figure 3.4F).

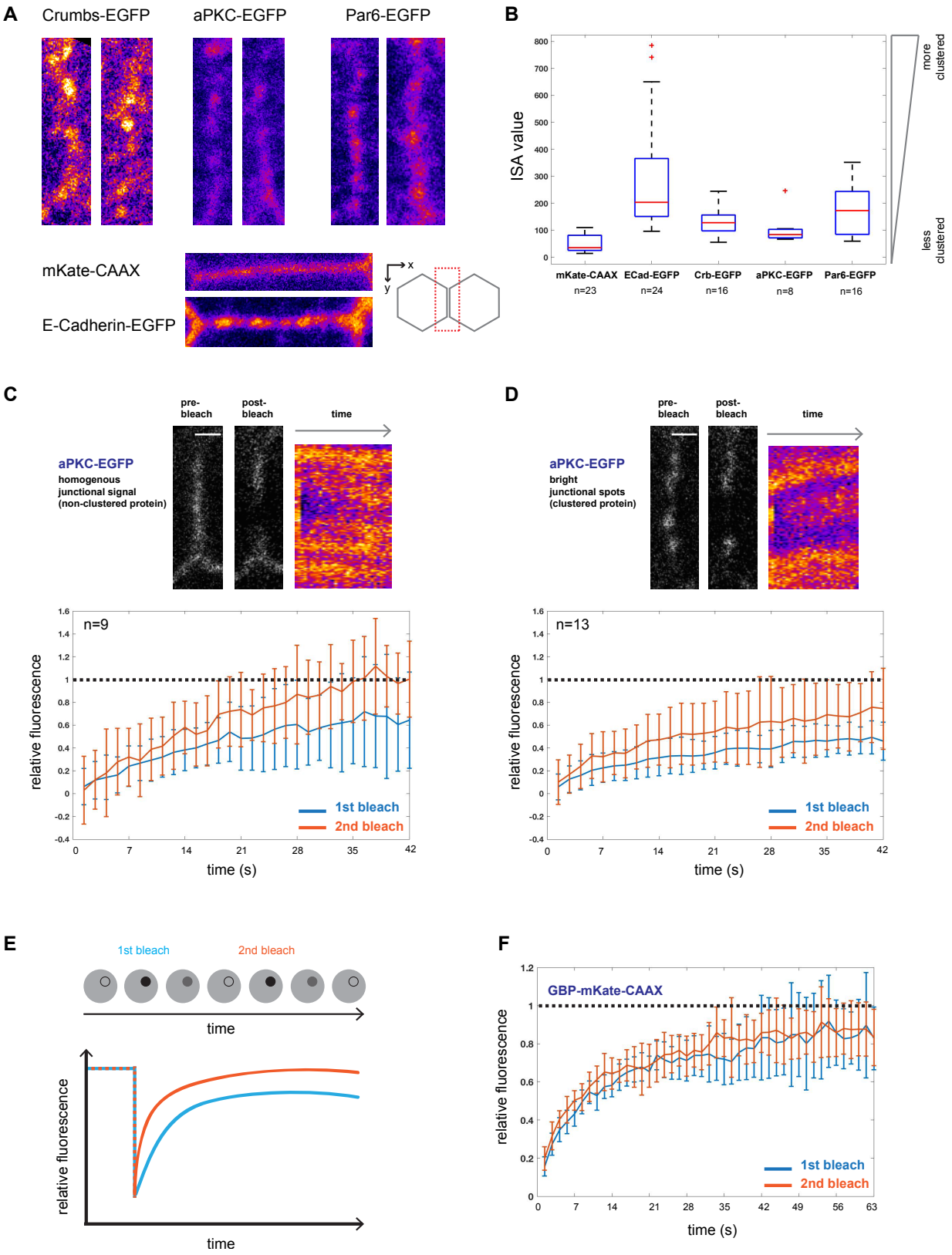


Figure 3.4 **Clustering of apical polarity proteins and mobility of aPKC.** Legend on next page.

Figure 3.4 *(previous page)* (A) Examples of confocal images showing the junctional distribution of Crumbs, aPKC, and Par6, respectively (endogenously-tagged with GFP), in live follicle epithelial cells. Additionally, mKate-CAAX and E-Cadherin-GFP are shown as examples of homogenous and non-homogenous distributions, respectively. Right bottom: a schematic showing top view of two neighbouring epithelial cells, the junction between them is boxed as an example of what is shown in confocal images. (B) Integrated signal anisotropy (ISA) along the junctions in confocal images. ISA was calculated by multiplying standard deviation of mean pixel value along the junction with the frequency of bright peaks along the junction. Higher ISA value means less homogenous fluorescent signal along the junction. (C) Top: An example of homogenous aPKC-GFP signal (non-clustered protein) along the junction before and just after photobleaching with the corresponding kymographs. Bottom: FRAP curves obtained. Second FRAP was performed 10-15 seconds after the first FRAP experiment. Mean ± SD. (D) Top: An example of bright junctional spot aPKC-GFP signal (clustered protein) along the junction before and just after photobleaching with the corresponding kymographs. Bottom: FRAP curves obtained. Second FRAP was performed 10-15 seconds after the first FRAP experiment. Mean ± SD. (E) Principle behind double bleaching FRAP experiment. After the first FRAP immobile fraction is bleached and the recovery after the second FRAP is complete since only the mobile fraction is bleached. (F) FRAP curves obtained in the control experiment, where GBP-mKate-CAAX was bleached that is homogenously distributed along the junction.

Next I performed the two-step FRAP on aPKC-GFP. The small regions in the junctions exhibiting homogeneous fluorescent signal (non-clustered signal) recovered to 100% of the initial signal before the second bleach (Figure 3.4C). However, bright fluorescent spots (clustered signal) recovered to only 70% of the initial signal before the second bleach (Figure 3.4D). This could be explained either by incomplete bleaching in the first step, or by slow exchange of part of the immobile fraction so that some signal recovered between the first and the second bleaching step also included some of it. This would suggest that the immobility persist on the order of minutes.

Altogether these results suggested that investigated apical polarity proteins form diffraction-limited clusters along the cell junction. In the case of aPKC appears, it appears that they are less mobile than non-clustered areas. These results were encouraging for further spatial analysis of the mesoscopic organisation of these proteins using super-resolution microscopy.

3.2.2 Endogenous Halo- and SNAP-tagging of polarity proteins

Super-resolution imaging approaches have traditionally used antibodies to label the intracellular structures. This method is ideal for qualitative visualising intracellular

proteins that are part of a known structure (e.g. microtubules, mitochondria, nuclear pores). However, this it is not the best approach for quantification of protein numbers, since multiple antibodies can bind to the epitope. Moreover, for numerous proteins (e.g. polarity proteins) antibodies do not exist or perform badly, especially for certain experimental systems (e.g. *Drosophila melanogaster*, *Caenorhabditis elegans*).

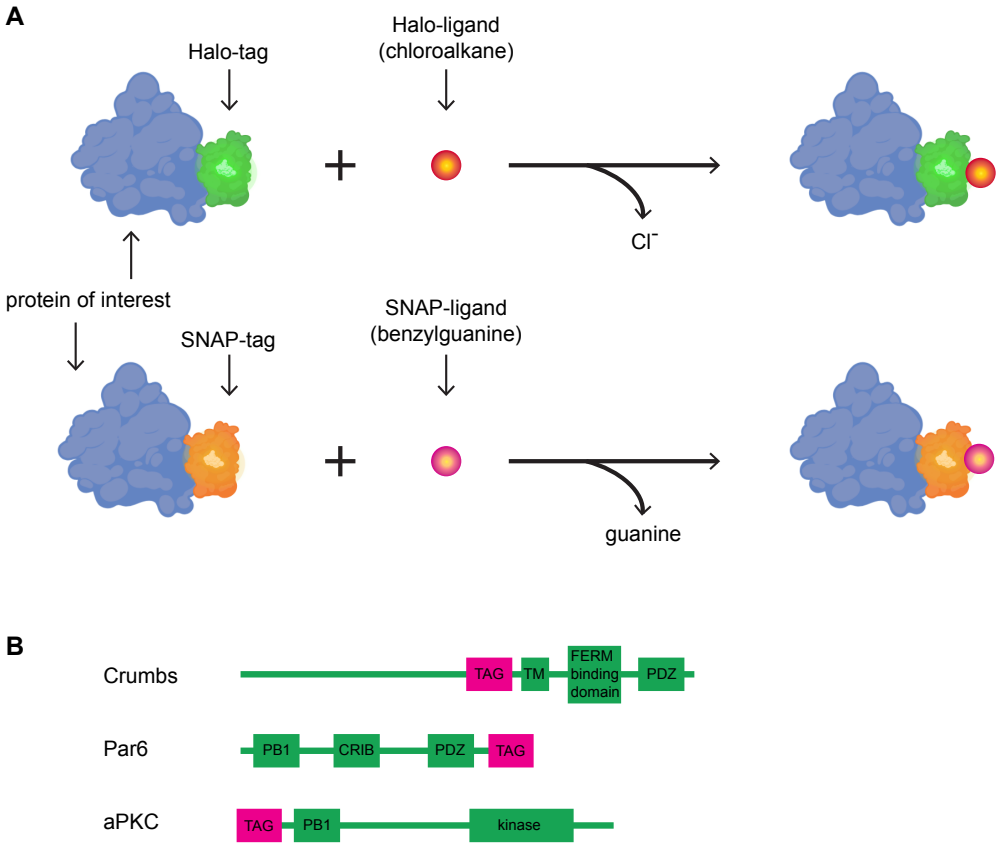


Figure 3.5 **Labelling of protein of interest using Halo- and SNAP-tagging system.** (A) Top: Principle of Halo-tag labelling. A protein of interest is fused to Halo-tag, which upon the presence of Halo-ligand (chloroalkane) reacts with it by forming a covalent bond. Chloride ion is released in this reaction. Bottom: Principle of SNAP-tag labelling. A protein of interest is fused to SNAP-tag, which upon the presence of SNAP-ligand (benzylguanine) reacts with it by forming a covalent bond. Guanine is released in this reaction. (B) Location of self-labelling tags in investigated polarity proteins. The genetic sequence of each tag was introduced to *Drosophila melanogaster* genomic sequence using CRISPR/Cas9 system.

To this end, I decided to utilise the Halo and SNAP labelling systems of genetically-encoded self-labelling proteins. The Halo tag is a modified haloalkane dehalogenase designed to covalently bind to synthetic chloroalkane (CA) ligands (Los et al. 2008). These ligands

can be attached to various different functionalities, like fluorescent dyes. Similarly, the SNAP tag is a modified version of the human DNA repair protein, alkylguanine-DNA alkyltransferase, that can react with benzylguanine (BG) derivative ligands (Keppler et al. 2003). Importantly, the stoichiometry of both reactions yields exactly one ligand molecule per protein tag, which enables exact quantification of the labelled proteins (Figure 3.5).

The CRISPR/Cas9-mediated homologous recombination was used to endogenously tag polarity proteins with either Halo or SNAP protein tags. The tag location was usually the same as for the GFP tag previously reported for each polarity protein. All lines generated were homozygous viable. This suggest that the function of polarity proteins was left intact. I assayed if tagging the polarity proteins with self-labelling enzymes perturbs their normal spatial organisation using confocal microscopy. As a control I used fly lines that have polarity proteins endogenously tagged with enhanced green fluorescent protein (GFP) and are used as standard fly lines in other studies. I labelled fixed follicle cells expressing Halo-tagged polarity proteins with JF646-CA. The JF646 signal matched that of the GFP-tagged version of the protein (Figure 3.3B and Figure 3.6). This suggests that in terms of spatial distribution investigated Halo-tagged polarity proteins behave the same as GFP-tagged proteins. Since cell polarity was not perturbed the functionality is conserved. I came to the same conclusion upon checking the SNAP-tagged proteins labelled with AlexaFluor647-BG (AF647-BG) (data not shown).

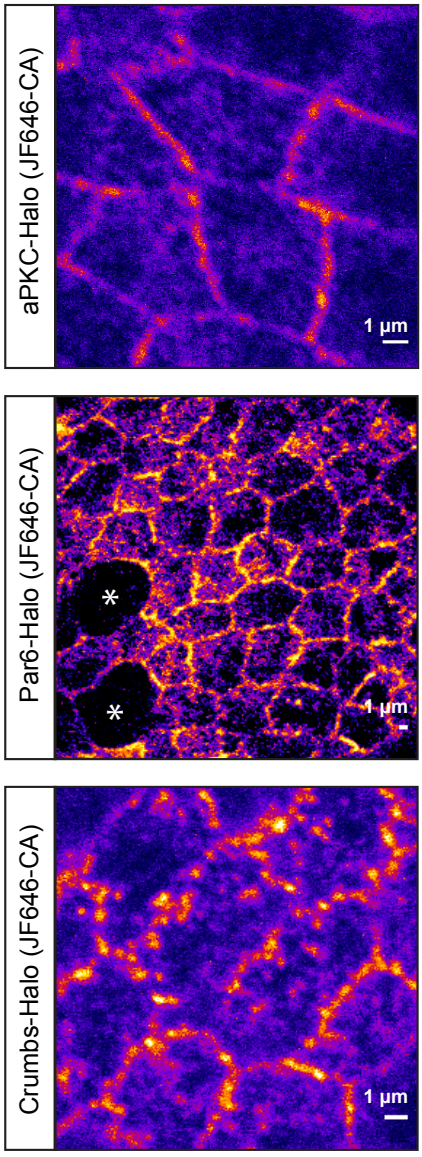


Figure 3.6 **Apical polarity proteins in follicle epithelial cells as seen with confocal microscopy.** Legend on next page.

Figure 3.6 *(previous page)* Imaging the marginal zone of the follicle cells expressing aPKC, Par6 and Crumbs that were endogenously tagged with the Halo-tag and labelled with JF646-CA. The asterisks denote the mitotic cells.

3.2.3 Preliminary experiments using dSTORM

Initially, this project started out by using wide-field dSTORM approach. I decided to first image the nuclear pore complex (NPC). The NPC consists mainly of nucleoporin proteins that are arranged in 8 subunits with a radial symmetry. In the cross section each subunit is built by proteins in the cytoplasmic ring (inserted in the outer nuclear membrane), inner ring (inserted between two nuclear membranes), and nucleoplasmic ring (inserted in the inner nuclear membrane) (Figure 3.7A) (Alber et al. 2007).

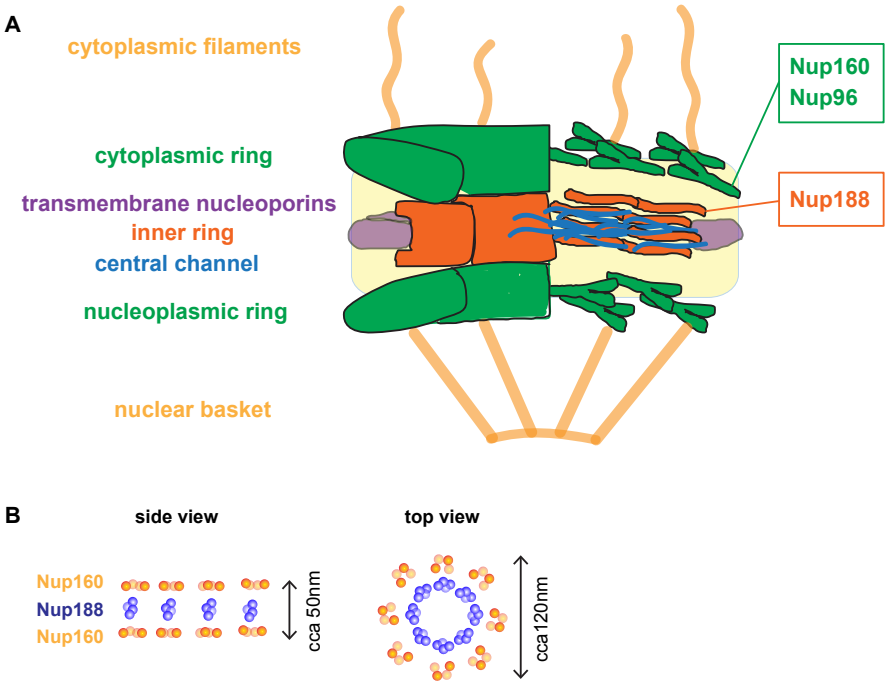


Figure 3.7 **General architecture of the nuclear pore complex (NPC).** (A) A simplified schematic of the NPC architecture with positional assignment of the nucleoporins used in this study. Redrawn from (Weberruss and Antonin 2016) (B) A schematic showing spatial arrangement of nucleoporin-160 (Nup160) and nucleoporin-188 (Nup188) within the nuclear pore complex.

Each subunit is composed of many different proteins called nucleoporins. Here, I used endogenously Halo- and SNAP-tagged Nucleoporin-160 (Nup160) that is present in the cytoplasmic and the nucleoplasmic ring. Moreover, I used endogenously SNAP-tagged

Nucleoporin-188 (Nup188) that is present in the inner ring. The stoichiometry of Nup160 and Nup188 is known with 4 protein copies present in each subunit (2 in each ring for Nup160) yielding 32 copies of Nup160 and Nup188, respectively, per NPC (Weberruss and Antonin 2016) (Figure 3.7B). Both Nup160-tagged lines generated were homozygous viable, while both Nup188-tagged lines were only heterozygous viable and did not produce homozygous progeny.

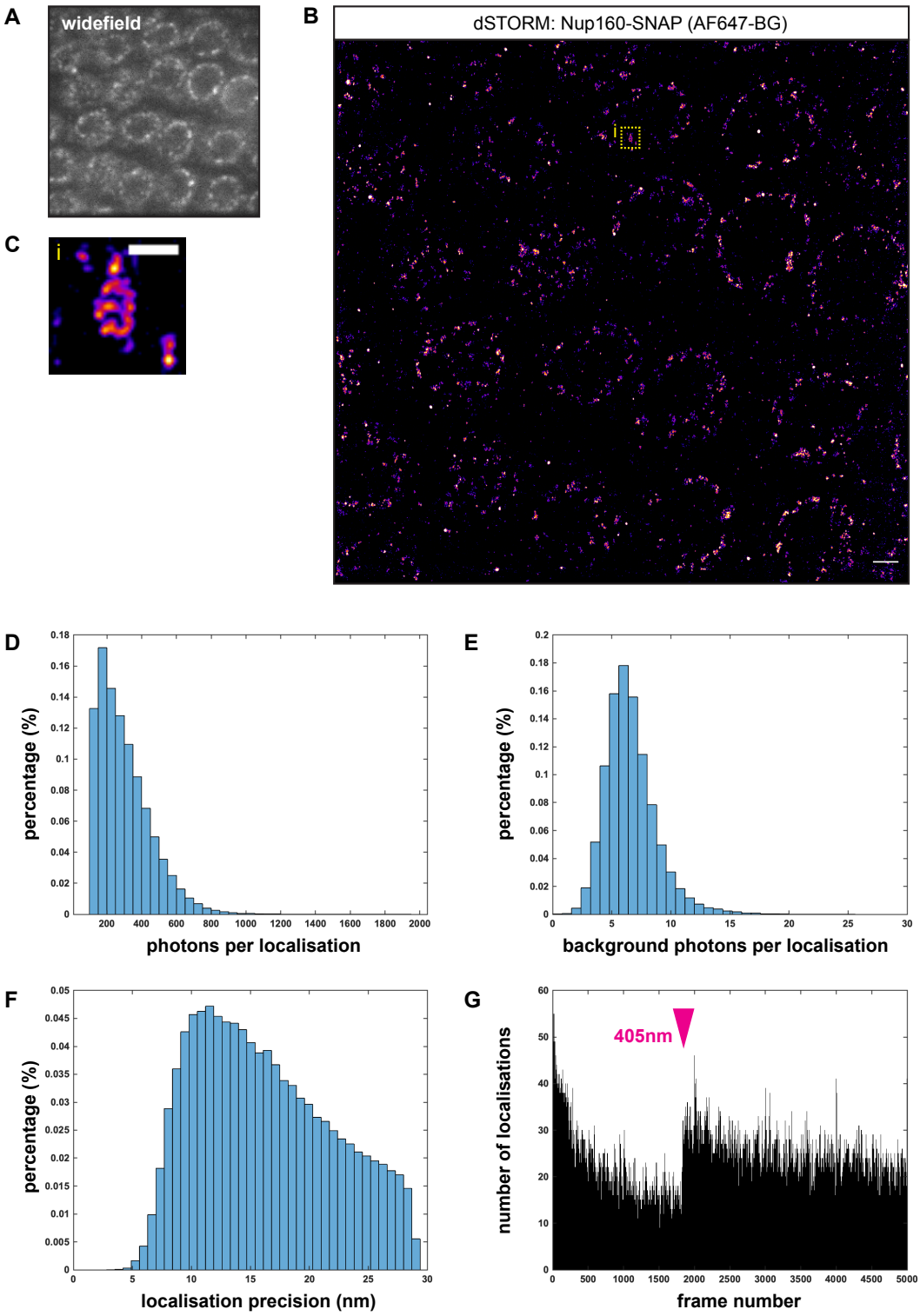


Figure 3.8 **Super-resolution imaging of Nup160-SNAP using dSTORM.** Legend on next page.

Figure 3.8 *(previous page)* (A) A widefield image of Nup160-SNAP labelled with AF647-BG in the nuclear membrane of the follicle cells. (B) A super-resolution image of Nup160-SNAP labelled with AF647-BG in the nuclear membrane of the follicle cells. The focal plane was positioned at the cross section of the nuclei. (C) A zoomed-in nuclear pore complex from the inset shown in (B). (D) Distribution of photons per localisation from the super-resolution acquisition shown in (B). (E) Distribution of background photons per localisation from the super-resolution acquisition shown in (B). (F) Distribution of average localisation precision per localisation from the super-resolution acquisition shown in (B). (G) Number of localisations per frame during super-resolution acquisition for the image shown in (B).

I labelled Nup160-SNAP in the follicle cells labelled with BG-AF647, that has been a standard dye used in dSTORM imaging (Dempsey et al. 2011). Based on the widefield image of the nuclei cross section the AF647 signal clearly appeared in the nuclear membrane (Figure 3.8A). However, acquisition of 5000 frames resulted in a super-resolution images of bad quality (Figure 3.8B), despite some NPC-like structures detected (Figure 3.8C). The number of localisations dropped very quickly from around 50 to 15 localisations per frame after first 1000 frames, as expected for AF647 (Figure 3.8G). The blinking was boosted by 405 nm wavelength laser illumination (Figure 3.8G). There was on average 326 photons per localisation (std=150), with 7 photons on average counted as background (std=2.1) (Figure 3.8D and Figure 3.8E). The localisation precision was on average 16 nm (std=5.8) (Figure 3.8F). Similarly, I could only observe NPC-like structures but not its details when the focal plane was set to the basal surface of the nuclei, despite a higher photon yield per localisation (1226 photons, std=31) and a higher localisation precision (9.2 nm, std=4.45) (Figure 3.9).

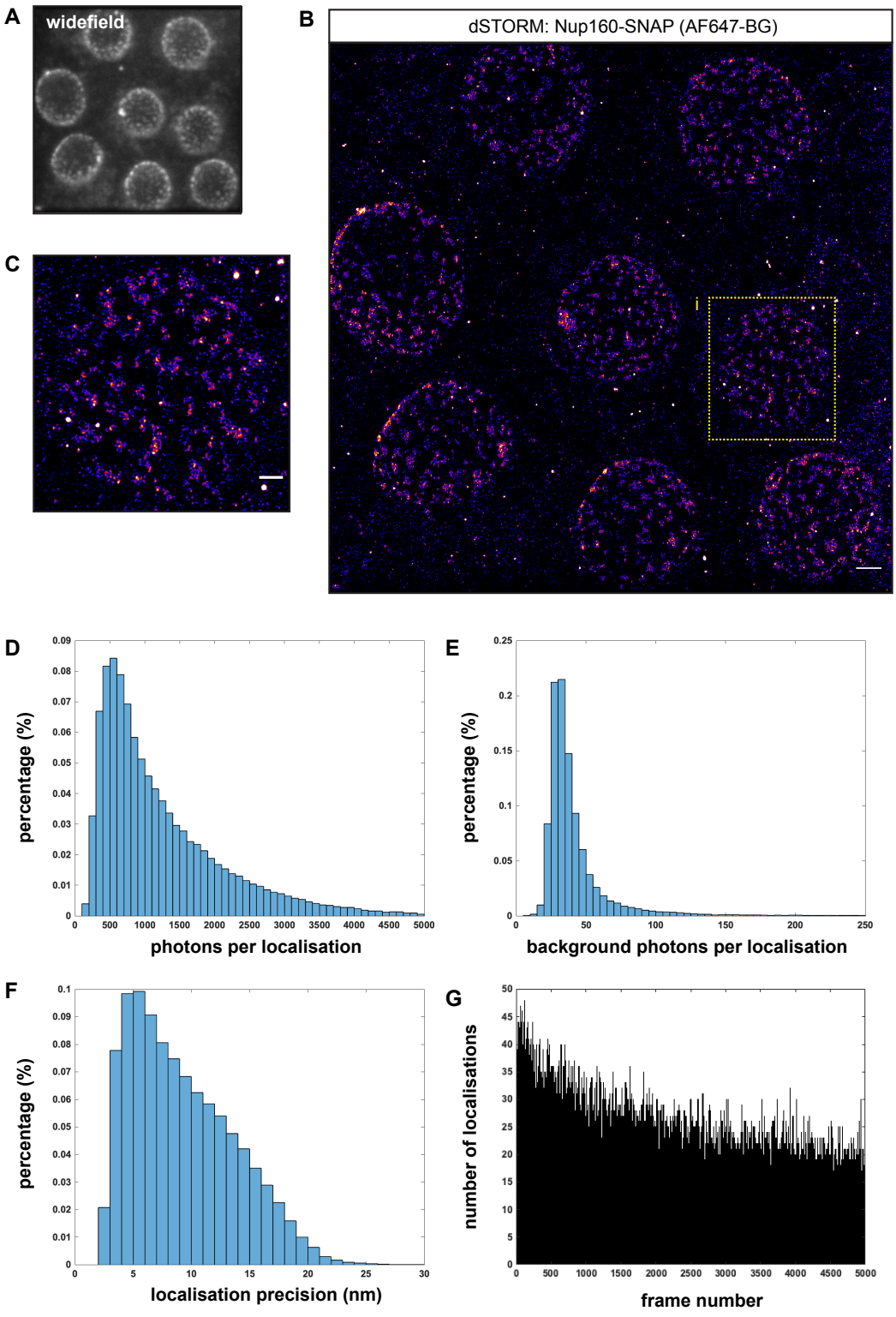


Figure 3.9 **Super-resolution imaging of Nup160-SNAP using dSTORM.** Legend on next page.

Figure 3.9 *(previous page)* (A) A widefield image of Nup160-SNAP labelled with AF647-BG in the nuclear membrane of the follicle cells. (B) A super-resolution image of Nup160-SNAP labelled with AF647-BG in the nuclear membrane of the follicle cells. The focal plane was positioned at the basal surface of the nuclei. (C) A zoomed-in nucleus from the inset shown in (B). (D) Distribution of photons per localisation from the super-resolution acquisition shown in (B). (E) Distribution of background photons per localisation from the super-resolution acquisition shown in (B). (F) Distribution of average localisation precision per localisation from the super-resolution acquisition shown in (B). (G) Number of localisations per frame during super-resolution acquisition for the image shown in (B).

In the context of possible molecule number quantification, I was wondering about the number of blinks per single AF647 dye molecule. For that I diluted AF647 on a coverslip to obtain dispersed single molecules of AF647 and quantified the number of blinks during the acquisition time. Under imaging settings used (see Chapter 2) a single molecule of AF647 produced on average 12.6 localisations (std=0.8) (Figure 3.10A) using dSTORM buffer. In a commercial buffer from Abbelight it was a bit lower and on average 8.8 localisations (std=0.8) (Figure 3.10B). Similar numbers were reported by other groups (Lin et al. 2015), however it is important to point out that the photo-switching behaviour is influenced not only by buffer but also by laser intensities, exposure time, etc. Importantly, this photo-switching behaviour is extremely stochastic and does not follow any temporal pattern.

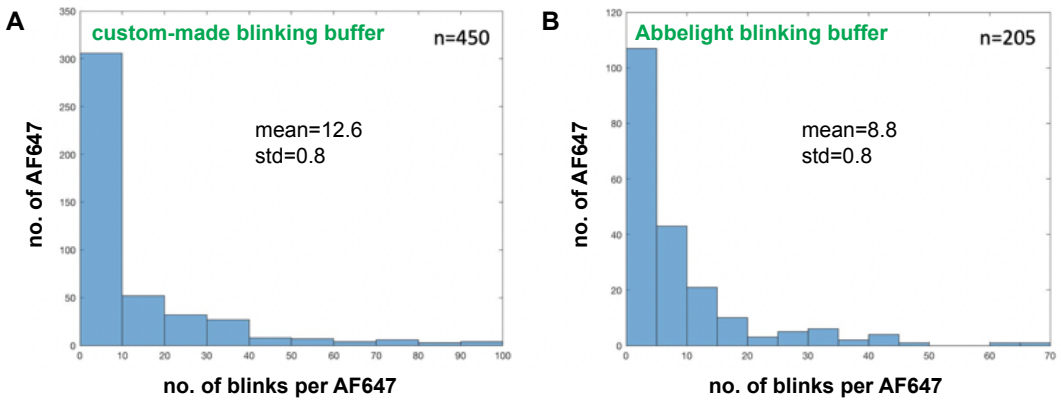


Figure 3.10 **Quantification of AF647 blinking in two different dSTORM buffers.** (A) Number of blinks per AF647 molecule in custom-made dSTORM buffer. (B) Number of blinks per AF647 molecule in commercial Abbelight buffer.

I next moved to image aPKC-Halo that I labelled with JaneliaFluor646-CA (JF646-CA). Acquired super-resolution images reconstructed from 20000 frames revealed clustered

localisations along the junctions, which supported observations made with the confocal microscopy (Figure 3.11A and Figure 3.11B). Interestingly, JF646 did not exhibit bleaching and the number of localisations per frame was constant (Figure 3.11F). There was on average 1200 photons per localisation, with 162 photons on average counted as background (Figure 3.11C and Figure 3.11D). The localisation precision was on average 16 nm (std=5.8) (Figure 3.11E).

While some dSTORM acquisitions were of good quality, most of them were not. I soon realised that my sample was drifting a lot in z-direction, hence only a small fraction of my image acquisitions did not feature the sample drift. Moreover, high background levels due to imaging deep into tissue resulted in small signal-to-noise ratio, which resulted in lower localisation precision (Figure 3.8F) than usually expected in the context of dSTORM imaging. Because of unreproducible quality of the acquisitions and stochasticity of the photo-switching behaviour, I reasoned that for quantitative super-resolution imaging dSTORM does not deliver the necessary conditions needed for counting the number of molecular targets.

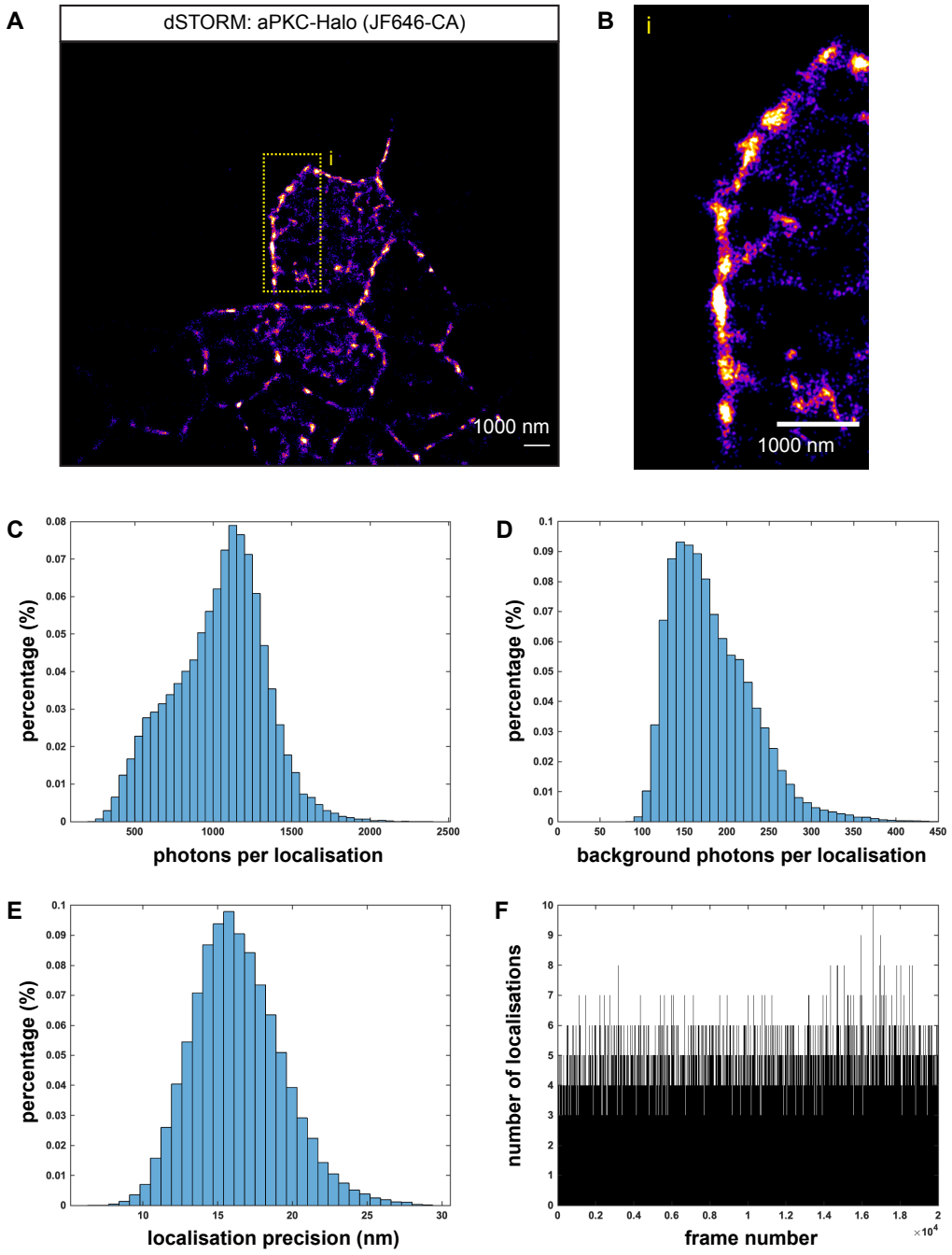


Figure 3.11 **Super-resolution imaging of aPKC-Halo using dSTORM.** Legend on next page.

Figure 3.11 *(previous page)* (A) A super-resolution image of aPKC-Halo labelled with JF646-CA in the follicular epithelium. (B) A zoomed-in junction from the inset shown in (A). (C) Distribution of photons per localisation from the super-resolution acquisition shown in (A). (D) Distribution of background photons per localisation from the super-resolution acquisition shown in (A). (E) Distribution of average localisation precision per localisation from the super-resolution acquisition shown in (A). (F) Number of localisations per frame during super-resolution acquisition for the image shown in (A).

3.2.4 DNA-PAINT

Because the dSTORM imaging approach did not bring satisfactory results, I switched to DNA-PAINT imaging after first year of my PhD studies. I decided to try this relatively new technique because it has a predictable behaviour of “blinks” and this is not influenced by the laser intensity or exposure time. Moreover, photobleaching does not influence sample imaging since the imager oligos are in the excess in the imaging solution and hence constantly replenished. Additionally, two technical optimisations on the optics side were applied to improve the quality of super-resolution acquisitions in work presented here. Firstly, wide field illumination was changed for slit scanning illumination. This means that the camera was operated in "light sheet" mode, introducing an effective confocal slit to reduce out-of-focus signal that originates along the z-axis and improve signal-to-noise ratio. Moreover, real-time drift correction in z-axis was implemented to the optical system (see Chapter 2).

3.3 Results

3.3.1 Imaging the nuclear pore complex

In order to validate DNA-PAINT imaging in combination with Halo and SNAP-labelling system, I decided to first image the nuclear pore complex (NPC). The NPC has a known symmetrical ring structure with 8 subunits as described above in Section 3.2.3.

First, I imaged Nup160-Halo using Cy3B-imager oligo in 4 nM concentration. I could observe single nuclear pores with various degree of labelled subunits suggesting incomplete labelling efficiency (Figure 3.12A-C). Importantly, the cross section view of the nuclei revealed two disks of localised signal (Figure 3.13A-C). This suggests that the observed structures are NPCs since the Nup160 is present in an outer and an inner nuclear membrane. Under optical setup conditions that I used the photon number per localisation was average 893 (std=328) (n=5 different egg chambers) and the number of background photons per localisations was on average 48 (std=12) (n=5 different egg chambers) (Figure 3.12D and Figure 3.12E). The calculated localisation precision was on average 9.11 nm (std=3.08) (n=5 different egg chambers) (Figure 3.12F). Importantly, photobleaching was less drastic than using dSTORM since the number of localisations per frame decreased less throughout the acquisition time (20000 frames) (Figure 3.12G). The number of photons per localisation and hence the localisation precision was the same when the focal plane was at the nuclear surface or at their cross-section (Figure 3.13).

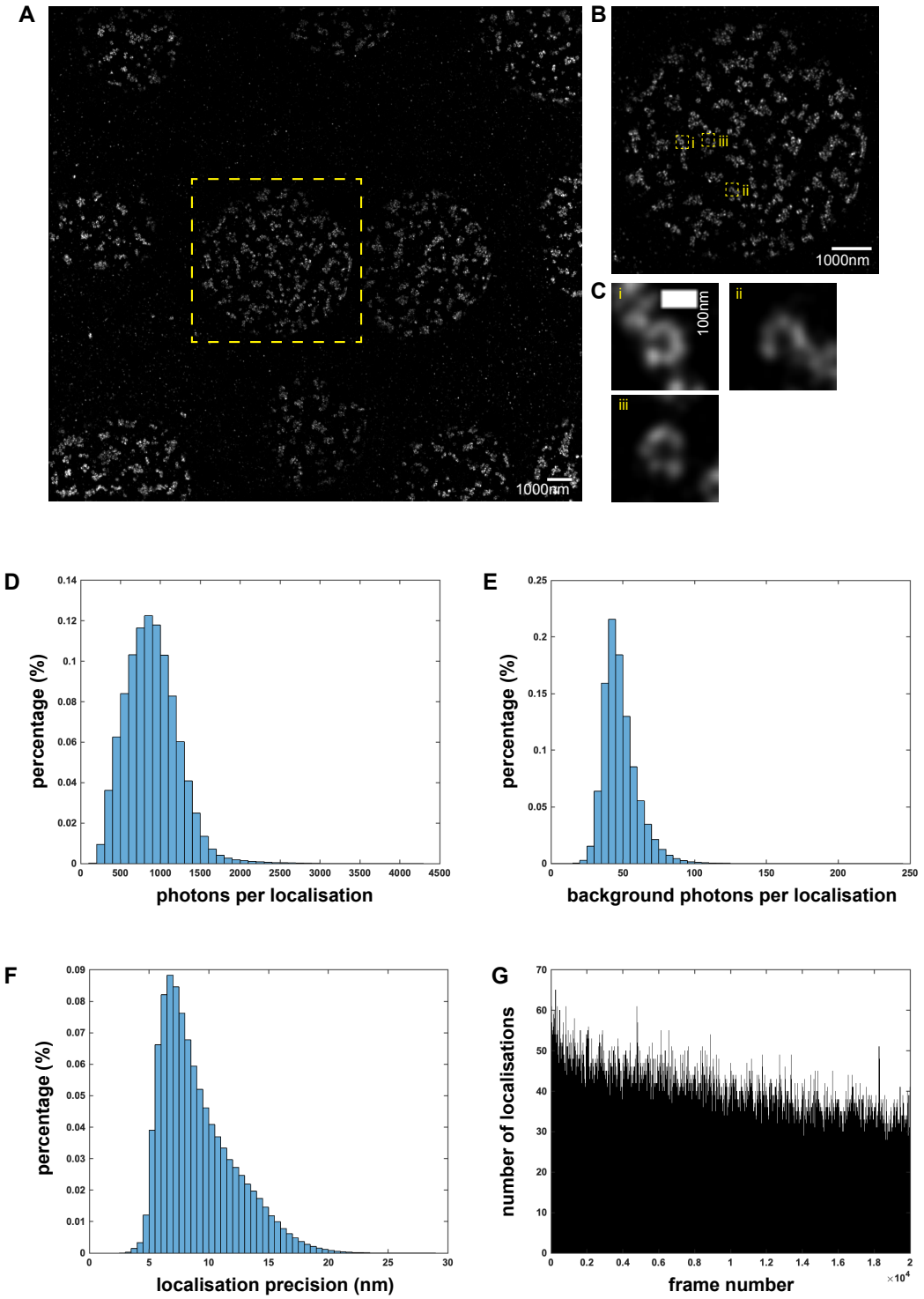


Figure 3.12 **Imaging nuclear pore complexes to validate DNA-PAINT.** Legend on next page.

Figure 3.12 *(previous page)* (A) A super-resolution image of Nup160-Halo in the nuclear membrane of the follicle cells using Cy3B-conjugated imager oligo (4 nM). The focal plane was positioned at the basal surface of the nuclei. Scale bar: 1000 nm. (B) A zoomed-in basal surface of the nucleus outlined with the dashed yellow box in (A). Scale bar: 1000 nm. (C). Three examples of the incomplete labelled nuclear pore complexes. Scale bar: 100 nm. (D) Distribution of photons per localisation from the super-resolution acquisition shown in (A). (E) Distribution of background photons per localisation from the super-resolution acquisition shown in (A). (F) Distribution of average localisation precision per localisation from the super-resolution acquisition shown in (A). (G) Number of localisations per frame during super-resolution acquisition for the image shown in (A).

Interestingly, I observed that nuclear pores were not evenly distributed, but were rather clustered. Since NPC were not reported to cluster in other experimental models used in super-resolution microscopy (i.e. U2OS cells), I wanted to check whether observation was caused by a fixation artefact. Therefore I performed live imaging of Nup160-Halo labelled with JF646-CA. Using the Zeiss system with the Airyscan, which has a theoretical resolution of up to 170 nm, I observed that the nuclear pores are de facto clustered in the fly follicular epithelium (Figure 3.14). This suggests that the fixation protocol used is not clustering the nuclear membrane structures.

Next, I examined whether it is possible to perform simultaneous two-colour imaging of Nup160-Halo and Nup188-SNAP using imager oligos conjugated to Cy3B and Atto655, respectively. In this case, two different orthogonal docking oligo sequences were used, here referred to as P1 and P3, respectively (the sequences are shown in Table 2.2 in Chapter 2). Correspondingly, the imager oligo sequences used for two-colour imaging were complementary to P1 and P3 docking oligo sequences. The images I obtained images showed the Nup188-SNAP signal sandwiched between two disks of Nup160-Halo signal, which corresponds to previously published structural data (Lin et al. 2019) (Figure 3.15A-B).

Under the optical setup conditions (see Chapter 2) used the calculated localisation precision was on average 9.11 nm (std=3.08) (n=5) and 10.7 nm (std=3.88) (n=3), for Cy3B and Atto655, respectively and the calculated photon number per localisation was on average 892 (std=328) (n=5) and 820 (std=426) (n=3) for Cy3B and Atto655, respectively. I observed that upon simultaneous double-laser (561 nm and 642 nm) illumination, the number of localisations per frame decreased more drastically for the Atto655 imager oligo (Figure 3.15D) than for the Cy3B imager oligo (Figure 3.15C). This was also reflected in

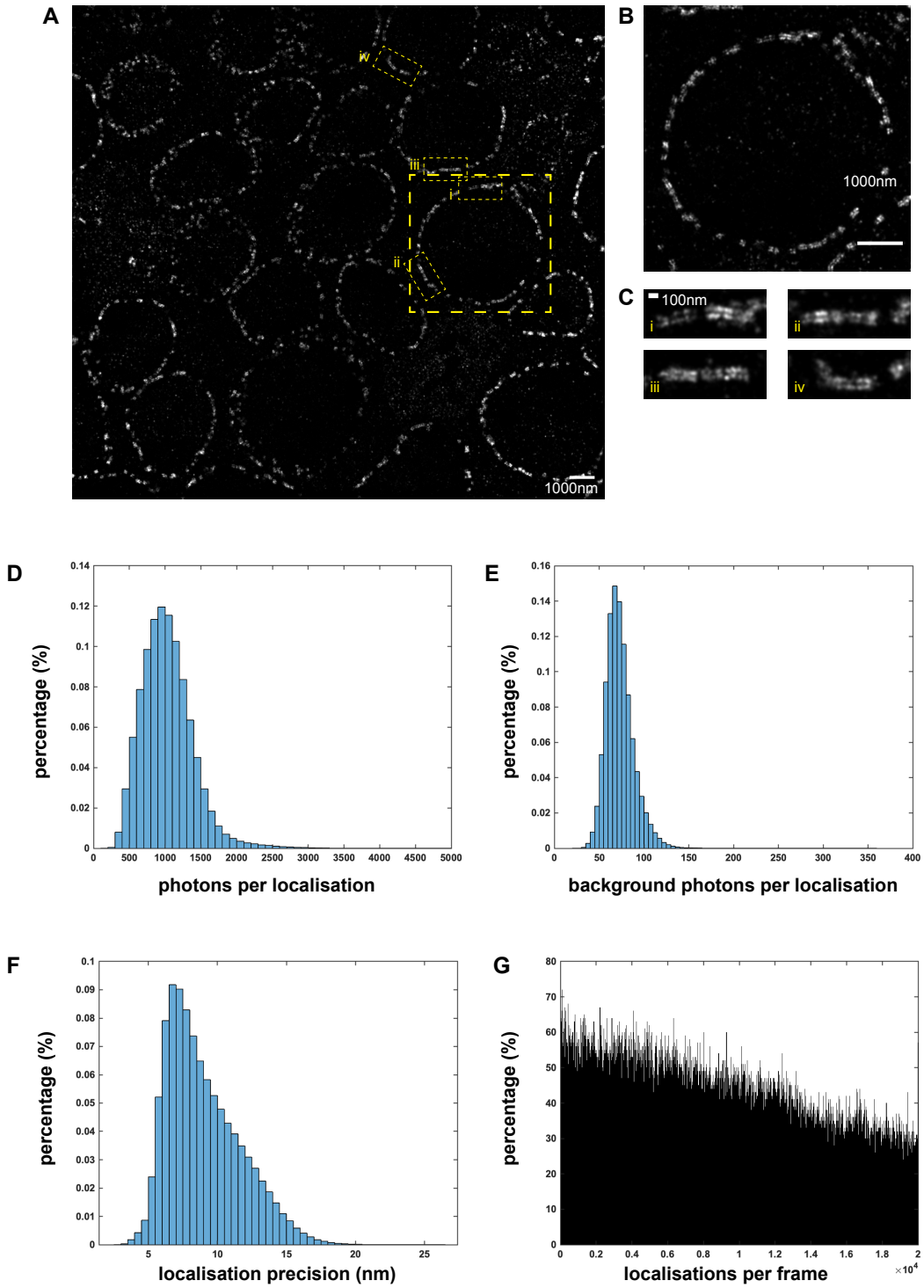


Figure 3.13 **Imaging nuclear pore complexes to validate DNA-PAINT.** Legend on next page.

Figure 3.13 *(previous page)* (A) A super-resolution image of Nup160-Halo in the nuclear membrane of the follicle cells using Cy3B-conjugated imager oligo (4 nM). The focal plane was positioned as a cross section, half-way through the nucleus in the centre. Scale bar: 1000 nm. (B) A zoomed-in cross section of the nucleus outlined with the dashed yellow box in (A). Scale bar: 1000 nm. (C) Four examples of the lateral views on the nuclear pore complex, where outer and inner rings are clearly distinguishable. Scale bar: 100 nm.(D) Distribution of photons per localisation from the super-resolution acquisition shown in (A). (E) Distribution of background photons per localisation from the super-resolution acquisition shown in (A). (F) Distribution of average localisation precision per localisation from the super-resolution acquisition shown in (A). (G) Number of localisations per frame during super-resolution acquisition for the image shown in (A).

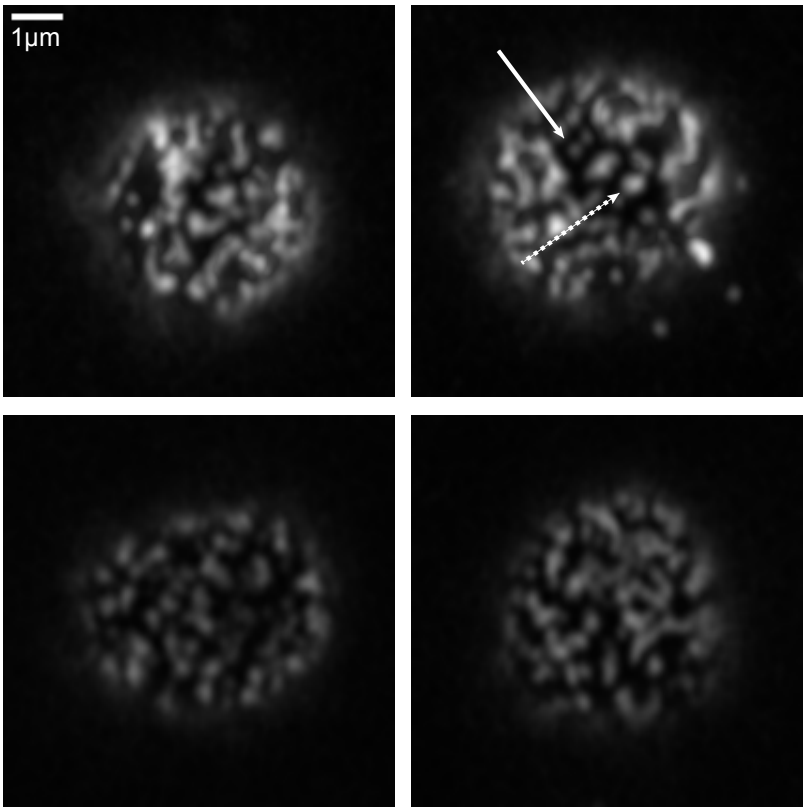


Figure 3.14 **Clustering of nuclear pore complexes in follicle cells.** Four examples of Nup160-Halo labelled with JF646-CA in live follicle cells. The images were acquired using confocal microscopy with Airyscan setup and focal plane was position at the basal surface of the nuclei. An arrow with a full line annotates a presumable single nuclear pore complex, while an arrow with a dotted line annotates a cluster of the nuclear pore complexes.

the overall number of fitted localisations: 644041 for Cy3B-conjugated imager oligo (52% of total detected localisations) and 151116 for Atto655-conjugated imager oligo (71% of total detected localisations).

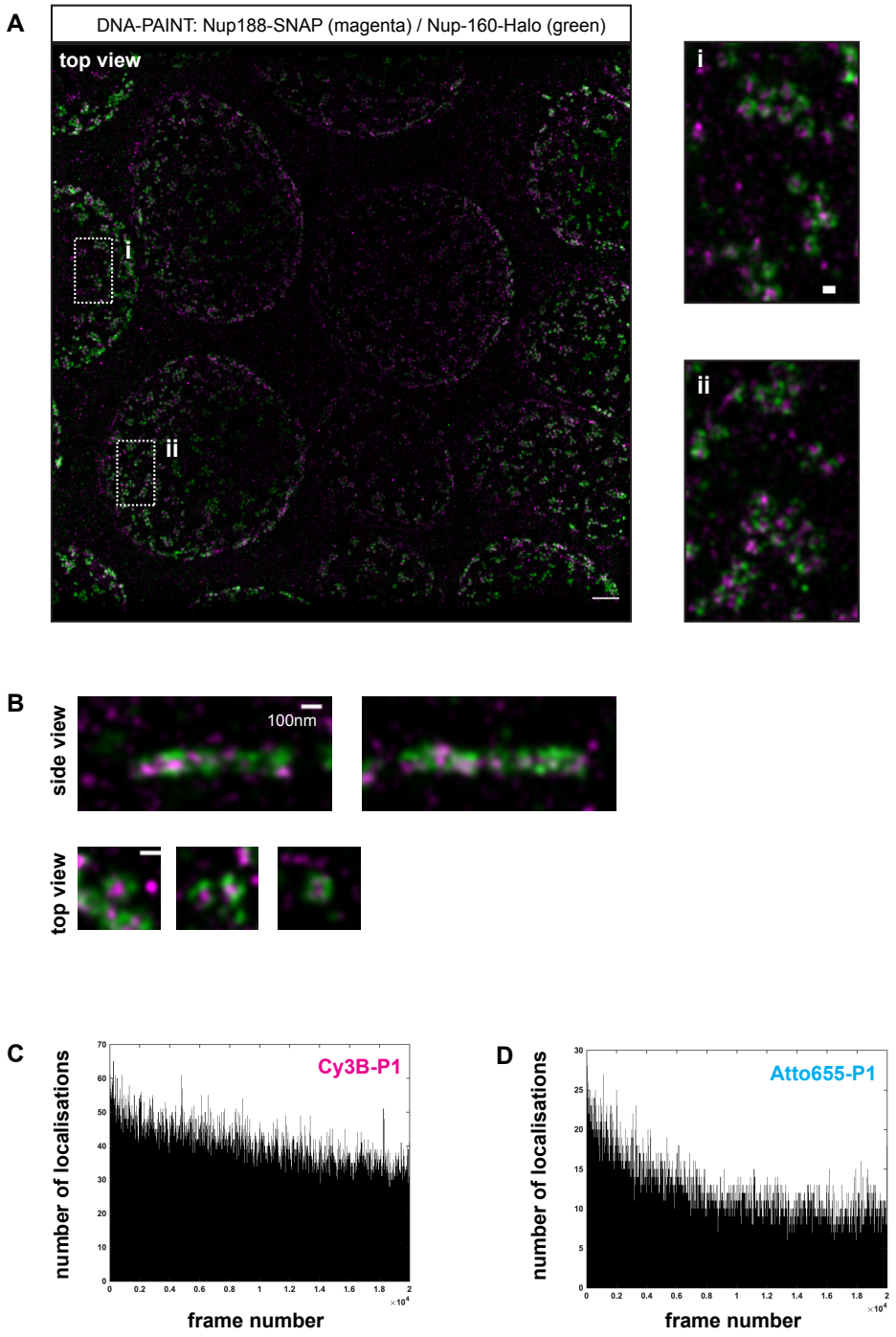


Figure 3.15 **Two-colour imaging of the nuclear pore complexes.** Legend on next page.

Figure 3.15 *(previous page)* (A) Super-resolution images of Nup188-SNAP (visualised using Atto655-P3 imager oligo) and Nup160-Halo (visualised using Cy3B-P1 imager oligo) in the nuclear membrane of the follicle cells. (B) Top: Two examples of Nup188-SNAP and Nup160-Halo signal imaged from the side. Bottom: Three examples of Nup188-SNAP and Nup160-Halo signal imaged from the top. (C) An example of distribution of the number of localisations per frame during super-resolution acquisition for Cy3B-P1 imager oligo. (D) An example of distribution of the number of localisations per frame during super-resolution acquisition for Atto655-P1 imager oligo.

These results suggest that upon simultaneous two-colour imaging 561 nm laser wavelength bleaches the far-red fluorophores and causes under-sampling of the structure.

3.3.2 Characterisation of the Cy3B and Atto655 bleaching rate

The bleaching of the Atto655-conjugated imager oligo made me wonder about the characteristics of the imager oligos that I was using upon illumination with each laser line.

For this, I imaged the follicular epithelium in an imager solution containing Cy3B- and Atto655-conjugated imager strand, respectively. This epithelium was not labelled with either Halo- or SNAP-ligands. Quantification of the number of the localisations per frame showed that the Cy3B-conjugated oligo exhibits an average between of 40 to 50 non-specific localisations per frame (Figure 3.16A and Figure 3.16C). For the Atto655-conjugated imager oligos this number was a bit lower and on average between 30 and 40 non-specific localisations per frame (Figure 3.16E and Figure 3.16G).

I also analysed how these non-specific localisations are spatially distributed within the imaging area. Interestingly, the temporal projection of all fitted localisations within 2D space for the Cy3B-conjugated imager oligo revealed that the central part of image contains considerably fewer less localisations than the peripheral parts (Figure 3.16B and Figure 3.16D). This was not the case for the Atto655-conjugated imager oligo, which exhibited a uniform number of localisations within the imaging area (Figure 3.16F and Figure 3.16H).

These results suggest that under the assumption that both imager oligos are diffusing within the tissue at the same rate, the Cy3B-conjugated imager oligo is more sensitive to photobleaching than the Atto655-conjugated imager oligo.

3.3.3 Quantification of the non-specific binding levels

Analysis of the imager oligos in the experiments described above showed that both imager oligos produce localisations as a result of non-specific binding. Therefore I wanted to characterize this "background" signal in more detail, since it could present a potential source of error in counting the target proteins.

Firstly, I created clones of wild-type cells that express a non-tagged version of polarity protein that cannot be labelled. These cells are juxtaposed in the tissue to cells expressing proteins with a self-labelling protein tag. This enables simultaneous imaging of both areas (Figure 3.17D) and bona fide quantification of the non-specific binding events. I imaged tissue that contained epithelial cells expressing aPKC-Halo protein and cells expressing a non-tagged (wild-type) version of aPKC with both Cy3B-conjugated and Atto655-conjugated imager oligos, respectively. Super-resolved reconstructed images revealed that non-specific binding events forms clusters that are randomly distributed and are not enriched at the junctions (Figure 3.17A and Figure 3.17B). In these images I quantified the total number of localizations in the cytoplasm of the non-tagged cells, in the cytoplasm of aPKC-Halo expressing cells and at the junctions of aPKC-Halo expressing cells and normalized the number of localizations to the area (see Chapter 2, Section 2.6.12). On average 44% (std=10, n=5) of the cytoplasmic localizations were due to non-specific binding events.

The background signal could arise from two sources: 1.) the remaining unreacted ligand molecules that were not washed out and then present docking sites for the imager oligo; 2.) from the non-specific binding of the imager strands. To distinguish between the two possibilities I imaged wild-type tissue that was not incubated with the docking oligo but had imager strand in the imaging solution. I then quantified the number of localizations per $1\,\mu m^2$ and compared it to the number of localizations in the wild type cells that were labelled with ligand conjugated to P1 docking strand. For Cy3B-P1 imager oligo there was on average 126 localizations per $1\,\mu m^2$ in the wild-type tissue that was not incubated with the docking strand, while on average 151 localizations per $1\,\mu m^2$ appeared in the tissue wild-type tissue that was incubated with the docking strand (Figure 3.17C). The difference was not statistically significant. For Atto655-conjugated imager oligo there was on average 280 localizations per $1\,\mu m^2$ in the wild-type tissue that was not incubated with the docking strand, while on average 298 localizations per $1\,\mu m^2$ appeared in the tissue wild-type tissue that was incubated with the docking strand (Figure 3.17C).

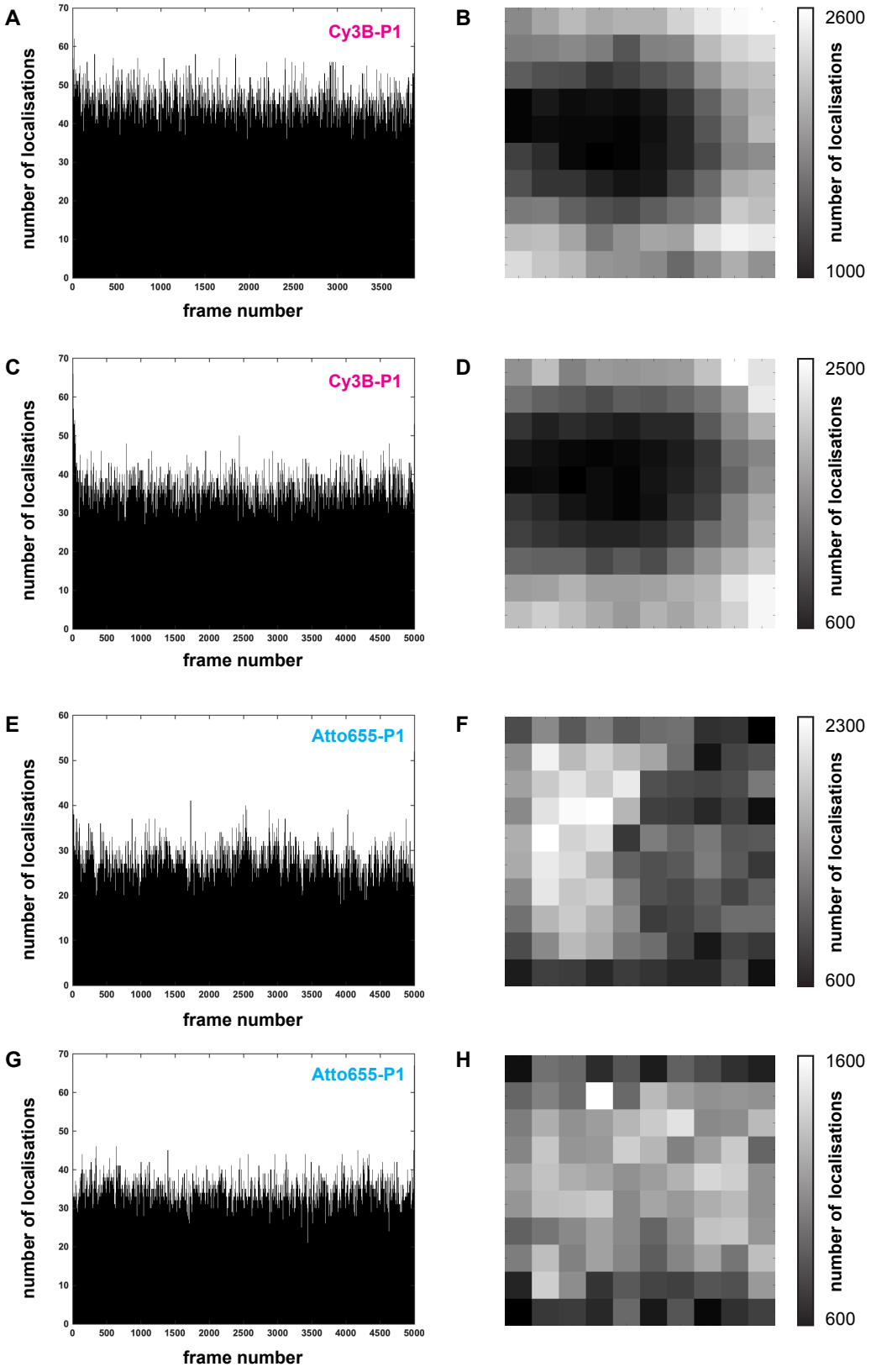


Figure 3.16 **Diffusion characteristics of Cy3B- and Atto655-conjugated imager oligos.** Legend on next page.

Figure 3.16 *(previous page)* (A and C) Two plots showing the number of localisations per frame in non-labelled follicular epithelial tissue using Cy3B- imager oligo (4 nM) from two different experiments. (B and D) Two maps showing the maximum projection of localisations shown spatially in the field of view (each square covers 4 µm^2), corresponding to plots shown in A and B. The number of localisations is colour coded with white corresponding to high number and black corresponding to low number of localisations. (E and G) Two plots showing the number of localisations per frame in non-labelled follicular epithelial tissue using Atto655-P1 imager oligo (4 nM) from two different experiments. (F and H) Two maps showing the maximum projection of localisations shown spatially in the field of view (each square covers 4 µm^2), corresponding to plots shown in E and G. The number of localisations is colour coded with white corresponding to high number and black corresponding to low number of localisations.

These results suggest that these “background" localisations come from non-specific binding (immobilisation) of the image oligo to some intracellular structures. Moreover, it appears that the Atto655-conjugated imager oligo is more “sticky” than the Cy3B-conjugated one.

To support this observation I then analysed the background signal in more detail. Visually inspecting the localisations confirmed that Atto655-conjugated imager oligo produced a more dense “background” footprint than Cy3B-conjugated one. However, both imager oligos produced clustered signal (Figure 3.18A and Figure 3.18B). I analysed the temporal trace of localisations for these clusters and I observed that the majority of the localisations are clustered in time (Figure 3.18C and Figure 3.18D). These clustered localisations suggest that the blinking appears to last several seconds. This is different from a temporal localisation trace for a NPC subunit, which is an example of real biological clustering (Figure 3.18E). The temporal trace of the NPC subunit area exhibits rather periodic appearance of the blinks, which is in line with the predictable behaviour of DNA hybridisation events (Jungmann et al. 2016). These observations support my hypothesis that non-specific localisations arise from immobilisation of the imager oligo to some intracellular structure until it gets bleached. Hence accumulated blinks result in a highly clustered signal in the super-resolution images.

3.3.4 Computational removal of the non-specific binding events

Next I wondered if localisations resulting from non-specific binding events could be removed in the post-processing before the image analysis. For this I used images of cell expressing aPKC-Halo juxtaposed to wild-type cells (clonal approach) (Figure 3.19A and Figure 3.20C). I reasoned that the average length of the bright and dark times for the non-specific

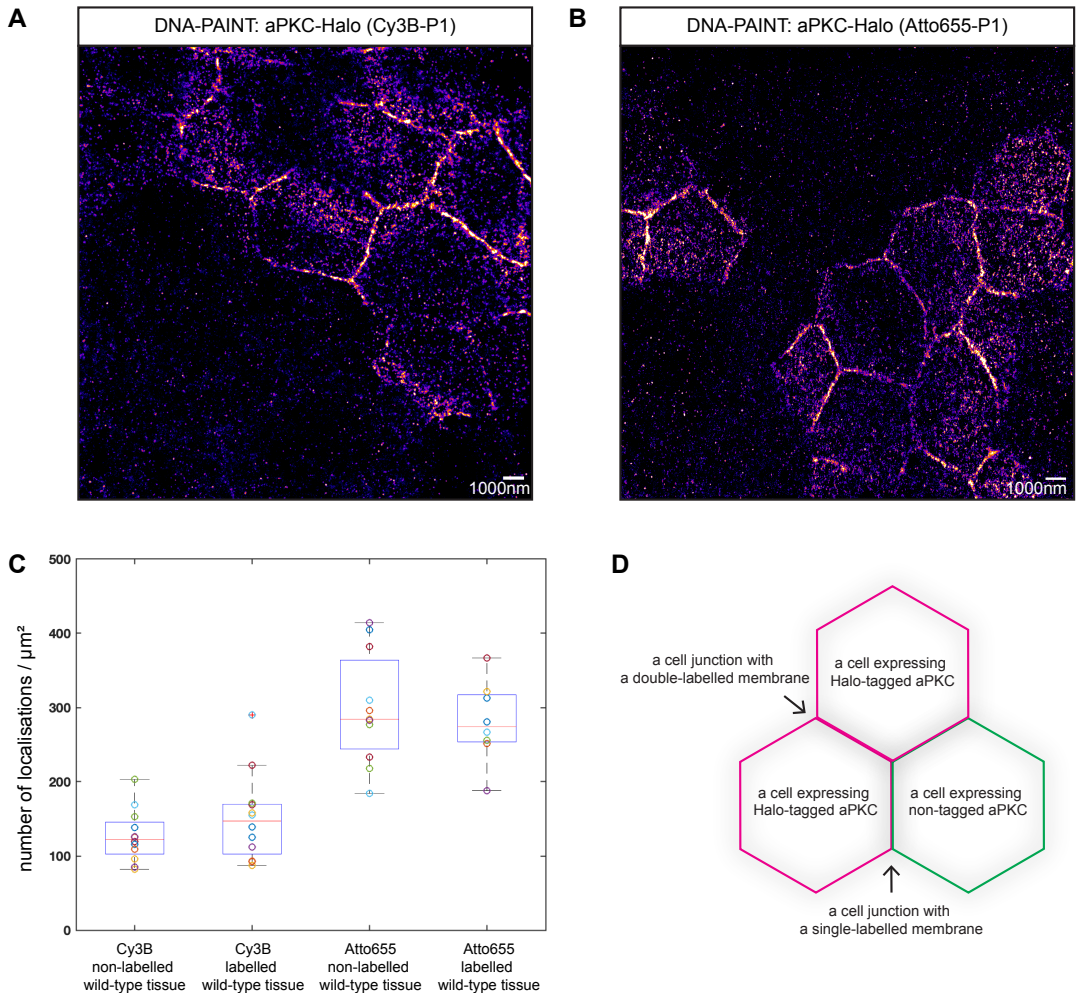


Figure 3.17 **Quantification of the Cy3B- and Atto655-conjugated imager oligos non-specific binding.** (A) A super-resolution image of aPKC-Halo (visualised using Cy3B-P1 imager oligo) in follicle cells. The labelled cells are juxtaposed to wild-type clone cells that are not expressing aPKC-Halo. (B) A super-resolution image of aPKC-SNAP (visualised using Atto655-P1 imager oligo) in follicle cells. The labelled cells are juxtaposed to wild-type clone cells that are not expressing aPKC-Halo. (C) Quantification of the number of localisations per 1 µm^2 in control (wild-type) tissue that was either labelled with a respective ligand or not following by visualisation with Cy3B-P1 and Atto655-P1 imager oligos, respectively. For each group cells in three different egg chambers were quantified. Two examples of with ligand labelled tissue (clone area) is shown in (A) and (B).

binding events should differ from those that come from the DNA hybridization-specific binding events. However, this was not the case and the populations overlapped (Figure 3.19B).

Next I analysed the standard deviation of localisation appearance per cluster over time. Deriving from the previous observation that temporal trace of "background clustered"

signal exhibits clustering of localisations in time (Figure 3.18), the temporal spread of localisations could be described with a standard deviation (with frame numbers of localisation appearances as an input).

Analysis of the distribution of the standard error values from the area of wild-type cells (only non-specific binding events) revealed that majority of them are distributed below 500 frames (low standard deviation value) (Figure 3.19C). On the other hand, analysis of the entire imaging area that included cells expressing aPKC-Halo resulted in a new peak around 1000 frames (Figure 3.19D).

Based on this I developed a two-step post-processing cleaning method for the raw data – that is a list of all fitted and localized blinks coming from both specific and non-specific binding events. This cleaning method first involves merging of localizations into clusters of around 15 nm in diameter. This I assumed based on localisation precision (around 10 nm), which would suggest that blinking from a single docking oligo creates a cluster of localisations with 10 nm diameter. This is experimentally confirmed in Chapter 5 (Section 5.3.1). This was followed by rejecting all single localizations (binding events that happened only once at the particular location) and clusters that have less than 10 of localizations (10 being minimum expected localizations events coming from a specific binding event in a given acquisition time frame). All remaining localizations were then used for further image analysis.

To illustrate this with aPKC-Halo as an example, I decided to reject all clusters that exhibited standard deviation of their time trace below 800 frames. This value was chosen since majority of clusters from non-specific binding events exhibit standard deviation value below 800. This resulted in 38% of clusters being rejected in the inside area and 69% of clusters rejected in the outside area (Figure 3.20A and Figure 3.20B). This fits with the previous calculations that around 44% of localisations (std=10) in aPKC-Halo images were due to non-specific binding events. Improving the rejection rate in the outside area seems challenging since the standard deviation values of cluster time-traces still overlap (Figure 3.20D and Figure 3.20E).

These results suggest that the differences in “blinking” behaviour between specific and non-specific binding events could be explored further in the future to computationally remove the background signal.

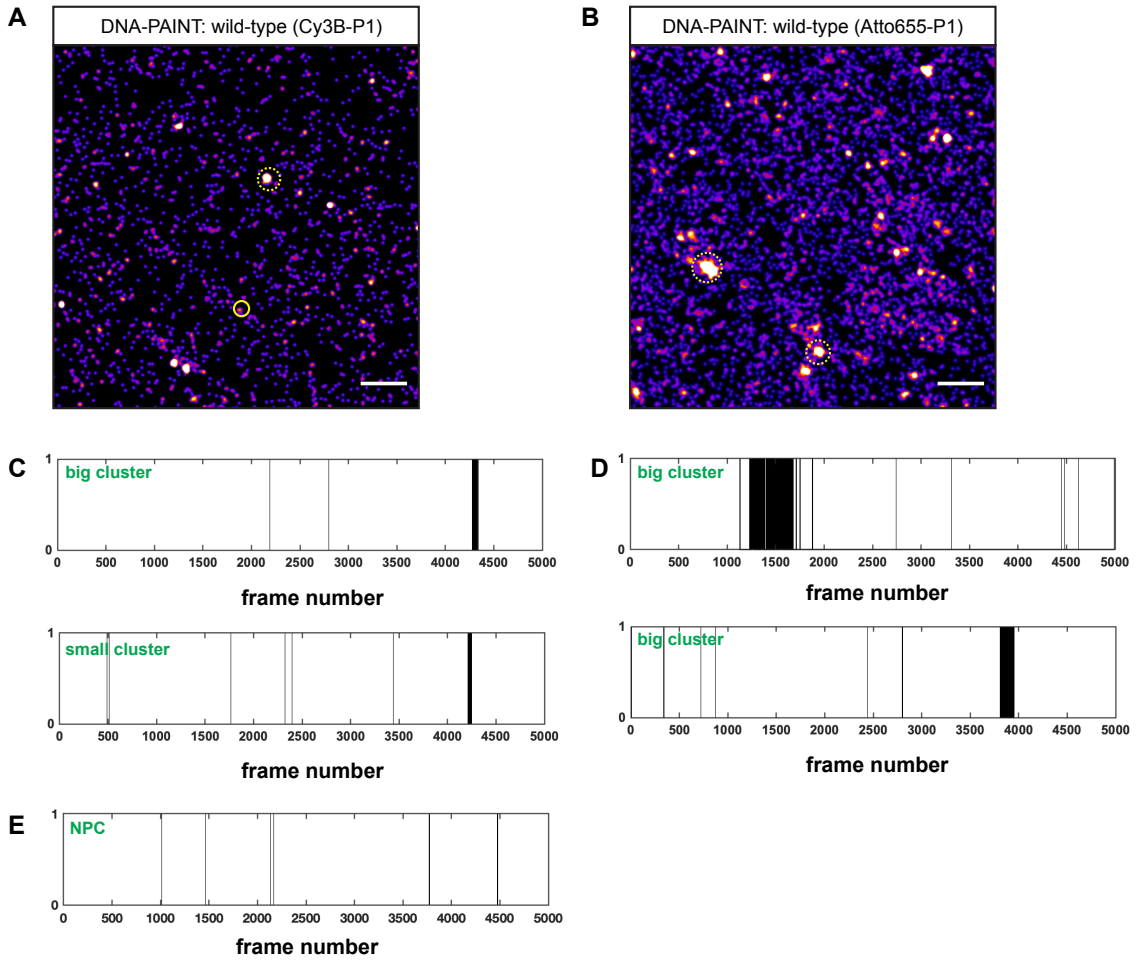


Figure 3.18 **Quantification of the imager oligos background footprint.** (A and B) An example of a super-resolution image of marginal zone of the wild-type cells using Cy3B-P1 or Atto655-P1 imager oligo (cells were not labelled with the docking strand beforehand). Dashed circles annotate big clusters of non-specific signal, while a full circle annotate a small cluster of non-specific signal. (C) An example of a temporal trace for a big and a small cluster of localisations for Cy3B-P1 imager oligo. A thin vertical line annotates a single localisation, multiple consecutive localisations appear as a wide vertical band (individual lines are not seen because of the low magnification). (D) An example of a temporal trace for a big and a small cluster of localisations for Atto655-P1 imager oligo. A thin vertical line annotates a single localisation, multiple consecutive localisations appear as a wide vertical band (individual lines are not seen because of the low magnification). (E) An example of a temporal trace for NPC subunit (a cluster of localisations arising from specific binding events) from (Figure 3.12C). The temporal distribution of localisations appears periodic indicating signal specificity.

3.4 Discussion

In this chapter I explored the distribution of apical polarity proteins in the fruit fly follicular epithelium by first using confocal fluorescence microscopy. Then I established a pipeline for super-resolution imaging of endogenously tagged proteins using the DNA-PAINT approach. As an imaging benchmark standard, the nuclear pore complex (NPC) was imaged and it was confirmed that two-colour imaging can be performed. Moreover, I characterised the background footprint of the imager oligos and proposed a computational method to remove the localisations arising from the non-specific binding events.

Virtually all super-resolution imaging studies conducted so far have used cultured cells as samples. There are a few examples when dSTORM was used in tissue samples (Woodhams et al. 2019; Heller et al. 2017; Herrmannsdörfer et al. 2017; Hu et al. 2016; Hou et al. 2014) and only a couple with DNA-PAINT in tissue samples (Jungmann et al. 2016; Park et al. 2018).

Technical limitations like light scattering and high background within the tissue sample are two of the most obvious reasons why the field has avoided imaging of thick biological samples. However, some biological processes, like epithelial polarity, cannot be studied in vitro. In this work electronic slit scanning was presented as one way to overcome the limitations mentioned above. Similarly, mechanical slit scanning was utilized to image mouse brain sections using DNA-PAINT (Park et al. 2018).

While the slit scanning approach improves the signal-to-noise ratio, the photon harvest is drastically reduced, due to effective exposure time being ten to fifteen times shorter. Nevertheless, considering that focal plane is tens of microns away from the coverslip and the imaging target is within the tissue, the mean photon yield per localisation was manageable. For comparison, a similar optical setup with line scanning confocal illumination and an exposure time of 300 ms yielded an average 3730 of photons per localisation (localisation precision on average 6 nm) using the Cy3B fluorophore in COS-7 cells (Park et al. 2018). However, an optical setup with spinning disk illumination could harvest over an order of magnitude more photons (Schueder et al. 2017). It is important to point out that the imaging buffers used in above mentioned studies had the same salt concentration and pH, the only difference was in the oxygen-scavenger system used (Trolox in their case).

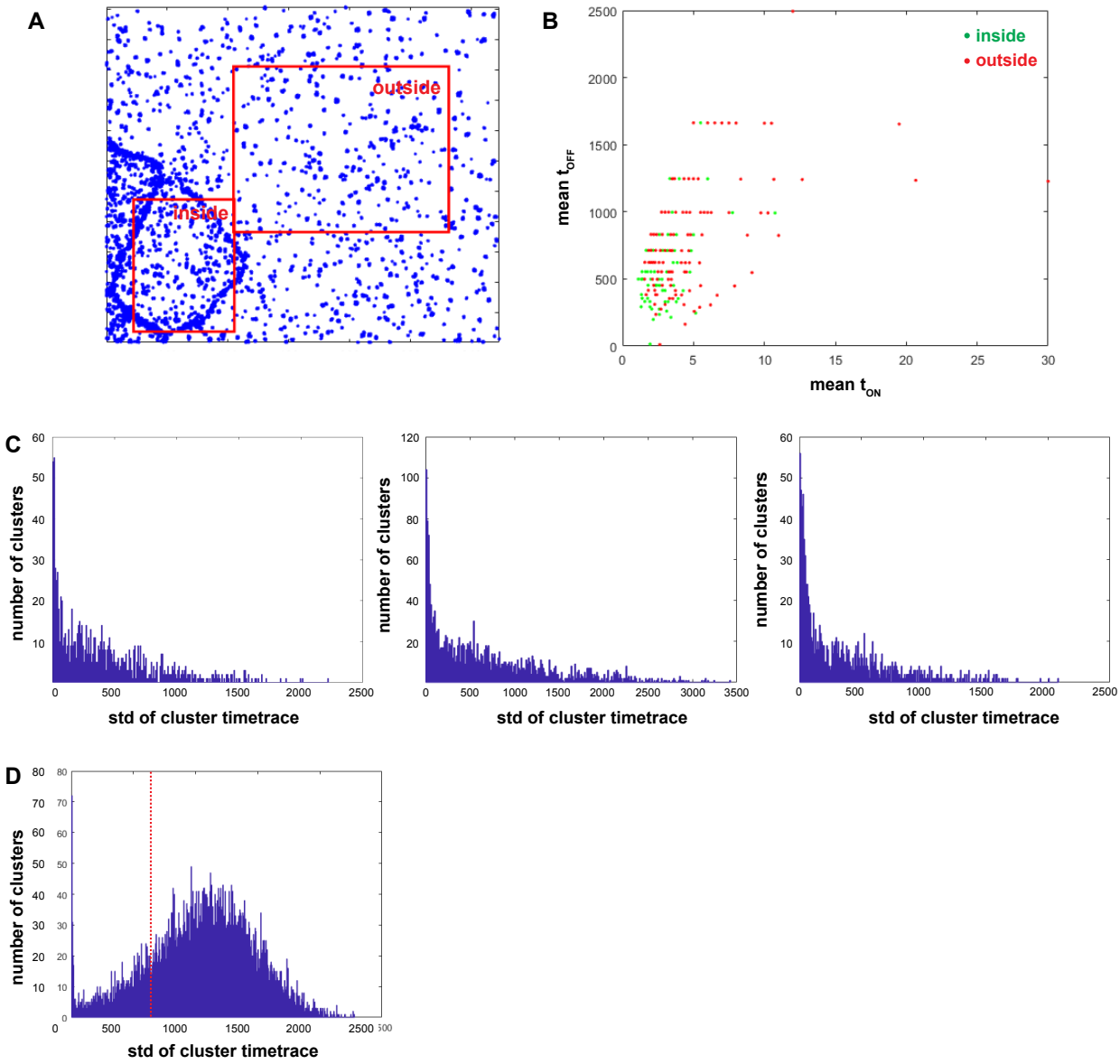


Figure 3.19 **Standard deviation of the localisation temporal trace as a parameter for removal of the nonspecific binding events.** (A) An example of segmented super-resolution image from (Figure 3.20C) showing clustered localisations with the "inside" area (follicle cells with labelled aPKC-Halo) and the "outside" area (wild-type follicle cells). (B) Correlation between mean t_{OFF} and mean t_{ON} for clusters in the inside and the outside area. (C) Distribution of standard deviation of cluster temporal trace for three different areas with non-specific localisations only. (D) An example of distribution of standard deviation of an area containing both specific and non-specific localisations.

The technical limitations are not only reflected on the side of the optical setup. In the case of the fruit fly egg chambers, sample preparation and mounting seems to be even more crucial. One of the key central points is sample immobilization. In the current sample mounting protocol, the older egg chambers are removed and only those in younger

stages are imaged. Despite using the minimum amount of the imaging solution, some egg chambers might still not be entirely immobilized between the cover slip and the glass slide and can exhibit the floating behaviour. This is especially pronounced at the beginning of imaging, when the sample is still "settling down". Upon imaging, the lasers are warming up the sample which might result in tissue expansion.

In my setup, the typical image acquisition time is 83 minutes (20000 frames, 250ms exposure time). Exposure times cannot be decreased since the photon harvest will decrease as well, and the imager strand concentration cannot be increased (to keep the sampling rate the same but decrease the imaging time) since the background will increase (again at the cost of the photon harvest).

One of the most challenging aspects of super-resolution microscopy is multi-colour imaging, due to the limited number of well-performing fluorophores (Dempsey et al. 2011). Here I demonstrated that simultaneous two-colour imaging is possible, but in our optical setup excitation of the sample with the 561 nm laser increases background and causes bleaching of the Atto655 imager oligos in the far-red channel. Despite the decrease in the localisation precision and sampling in the far-red channel, two-colour images can be still acquired. However, bleaching of the imager oligo strands presents a problem for quantitative imaging since the binding kinetics will not be traceable.

The original DNA-PAINT approach offered an elegant solution to this problem where different docking oligo species can be sequentially visualized using the same fluorophore (e.g. Cy3B) by washing away the imager solution with one imager strand before introducing a new one (Jungmann et al. 2014). However, this is possible when imaging samples in vitro, where highly permeabilised isolated cells that are attached to the cover slip allow thorough washing and replacement of the imager solution without physical perturbing the cell's position. For non-attached tissue samples, like the fruit fly egg chambers, this is not possible at the moment and would require designing a sample-specific micro-fluidic device. Moreover, how well can be the imager solution washed away and replaced in a multi-layered sample is another experimental challenge. Therefore it seems that the first attempt at solving this problem would be trying to correct things at the optical side and minimize the leakage of the fluorophore emissions due to 561 nm laser excitation into the far-red channel or trying to find a new combination of fluorophores, where channel cross-talk does not occur.

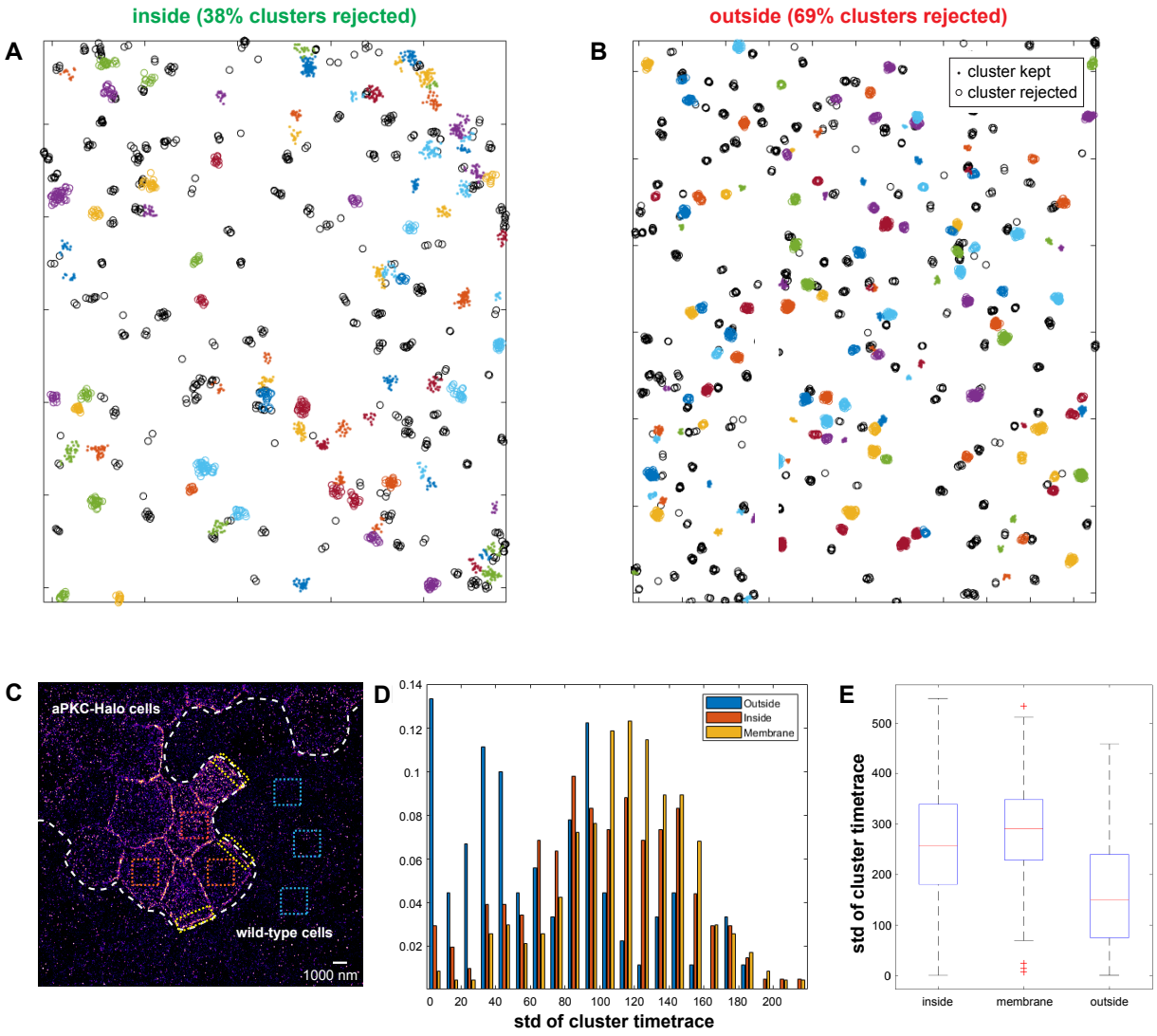


Figure 3.20 **Standard deviation of the localisation temporal trace as a parameter for removal of the nonspecific binding events.** (A and B) Removal of the clusters in the inside and the outside area based on standard deviation of their temporal trace (here std=800). Dots: clusters kept. Black circles: clusters too small. Coloured circles: std to small. (C) An example of a super-resolution image showing aPKC-Halo cells surrounded by a clone of wild-type cells. (D) Distribution of standard deviation of cluster temporal trace for three different areas in (C): outside (wild-type cells), inside (aPKC-Halo cells), membrane area (clone border). (E) Boxplots for standard deviation values of cluster temporal traces for three different areas in (C): outside (wild-type cells), inside (aPKC-Halo cells), membrane area (clone border).

Most recently, a new approach to multi-colour imaging in DNA-PAINT was introduced based on the duration and frequency of blinks. Again, the same fluorophore is used, but here the complementarity of the docking oligos is different for different targets (thus giving binding events of different durations). Additionally, the number of docking domains

introduces the change in the frequency of binding events that can be used to increase the number of targets (Wade et al. 2019). Importantly, when using the frequency as a variable the quantitative imaging would not be possible anymore since this is a parameter to count the number of molecular targets (Jungmann et al. 2016). However, the introduction of the two different lengths for complementary oligos would enable quantification. One has to keep in mind that the difference in blink duration would be substantial, which would be probably reflected in localisation precision through the photon yield ($k_{OFF} = 1.6s^{-1}$ for a 9 base pair complementarity and $k_{OFF} = 0.2s^{-1}$ for a 10 base pair complementarity) (Jungmann et al. 2010).

One particular problem of DNA-PAINT imaging is non-specific binding of the imager oligo, which results in non-specific localisations in the super-resolution image. Here I measure for the first time the levels of the non-specific binding. Importantly, the background comes mainly from the non-specific binding of the probe and not from specific binding of the probe to the docking strand that was not washed away during labelling protocol. Additionally, the signal-to-noise ratio decreases with an increase in the imager oligo concentration, and usually the unbound fraction becomes indistinguishable when concentration exceeds a few tens of nanomolar (Acuna et al. 2012). This problem could be solved by using DNA-PAINT imaging with fluorescence resonance energy transfer (FRET) probes (Deußner-Helfmann et al. 2018; Auer et al. 2017). Here the docking oligos are conjugated with an acceptor fluorophore (e.g. Atto 647N), while transiently binding imager oligos are conjugated with a donor fluorophore. (e.g. Atto 488). During imaging only the donor fluorophores are excited, and upon binding to a docking oligo, energy from the excited donor fluorophore is transferred to the acceptor fluorophore, whose emission is then detected downstream. However, the complexity of FRET-based DNA-PAINT imaging comes at the cost of performing quantitative imaging. Moreover, two channels are used to image a single colour, which would make two-colour imaging more challenging.

A more elegant approach in decreasing background fluorescence would be to use molecular beacons. These probes are hairpin-shaped single-stranded oligonucleotides with a fluorophore conjugated on one end and a quencher on the other (Tyagi and Kramer 1996). Upon hybridisation with the complementary sequence the quencher will extend away from the fluorophore and allow fluorescent emission. In this way the free floating imager oligos in DNA-PAINT would be quenched (personal communication, Kenny Chung, Bewersdorf laboratory). This would not only decrease the background fluorescence (and thus increase

the resolution), but would also allow image acquisition at a higher frame rate, since free floating imager oligos would not be detected. Importantly, fluorescence from non-specific binding events would also not be detected.

Finally, in order to be able to make DNA-PAINT acquisitions competent for quantification of molecular target number it is crucial to know the stoichiometry of the labelling probe to your molecular target. Antibody labelling, which was used in the original DNA-PAINT approach is not suitable. Here I utilize Halo and SNAP ligands that bind to respective self-labelling tags with a 1:1 stoichiometry. This approach was recently used to resolve nuclear pore structure in cultured cells (Schlichthaerle et al. 2019; Thevathasan et al. 2019). Interestingly, they used a different molecular linker to conjugate the Halo ligand to the docking strand in order to keep the ligand reactivity intact. In their approach, they also used a standard Halo ligand (commercially available from Promega), which in my hands lost reactivity upon conjugating with the oligonucleotide. It would be interesting to see if the Halo ligand that I used and the one used in recently published studies (Schlichthaerle et al. 2019; Thevathasan et al. 2019) differ in the labelling efficiency, which I will address in the next chapter.

3.5 Perspectives

Overall, the results in this chapter demonstrated the ability of the DNA-PAINT approach to resolve the mesoscopic features of protein organisation within a thick tissue sample. The example of it presented in this work was the fruit fly egg chamber. Despite the fact that I demonstrated simultaneous two-colour imaging using Cy3B- and Atto655-conjugated imager oligos, there were some technical limitations that would significantly influence the quantification of two-colour super-resolution images. Therefore, I decided to focus on quantitative single-colour super-resolution imaging. In the following chapter, I describe the principle behind quantifying the number of molecular targets in the super-resolution images acquired with the DNA-PAINT approach. I experimentally test counting using DNA-origami. I then explain how I used the nuclear pore complex in the fruit fly follicle cells to calibrate the influx of the imager strand necessary for quantifications. Moreover, I also investigate labelling efficiency of Halo- and SNAP-tagged proteins.

3.6 Acknowledgment of contributions

The quantification of the integrated signal anisotropy was developed in collaboration with Richard Butler from the Imaging Facility of the Gurdon Institute. The FRAP analysis pipeline was developed by Julia-Falo Sanjuan. The endogenous-tagging of polarity proteins was carried out by Nick Lowe. The endogenous-tagging of Nup160 and Nup188 was carried out by Jenny Richens and Amandine Palandri. For the post-processing of super-resolution images for the background removal, I collaborated with Leila Muresan at the Cambridge Advanced Imaging Centre.

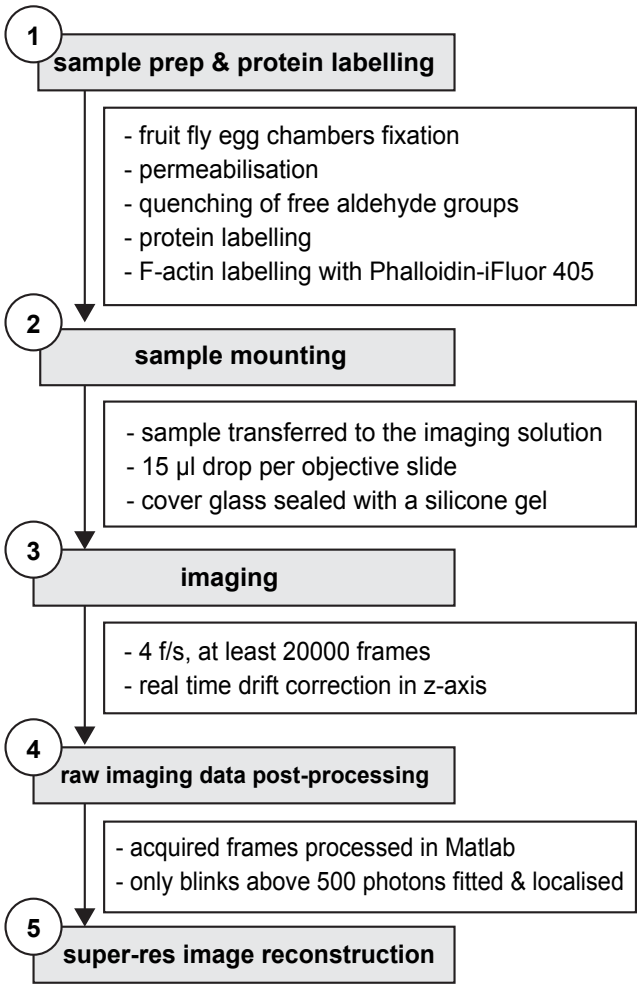


Figure 3.21 **Work-flow for super-resolution imaging of proteins in fruit fly egg chambers.**

Chapter 4

Calibrating DNA-PAINT for protein counting *in vivo*

4.1 Introduction

As detailed in Chapter 3, super resolution methods allow the visualization of biological structures at the nanoscale. However, despite up to an order of magnitude improvement in spatial resolution counting individual proteins in densely packed areas is limited by poorly understood photophysical behaviour of organic fluorophores (Vogelsang et al. 2010). DNA point accumulation imaging in nanoscale topology (DNA-PAINT) is an alternative method that utilises DNA hybridisation to achieve single molecule blinking behaviour and, subsequently, super-resolution images (Jungmann et al. 2014). Importantly, DNA hybridisation is predictable and can be described using a simple kinetic model thus enabling protein counting (Jungmann et al. 2016). Here, I describe the use of DNA-PAINT for counting the proteins in the fruit fly follicular epithelium. I invite the reader to refer to the experimental work-flow to facilitate reading of this chapter (Figure 4.1).

4.1.1 Overview of quantitative DNA-PAINT

Quantification of the number of binding sites using DNA-PAINT or quantitative DNA-PAINT (qPAINT) was originally described by Jungmann and colleagues (Jungmann et al. 2016). qPAINT utilizes predictable binding kinetics between single strand sequences of nucleic acids upon hybridisation with a complementary strand (Figure 4.2A). Single DNA strand hybridisation (meaning association) and dissociation can be described with a kinetic model that follows a second-order reaction rate upon association and a first-order

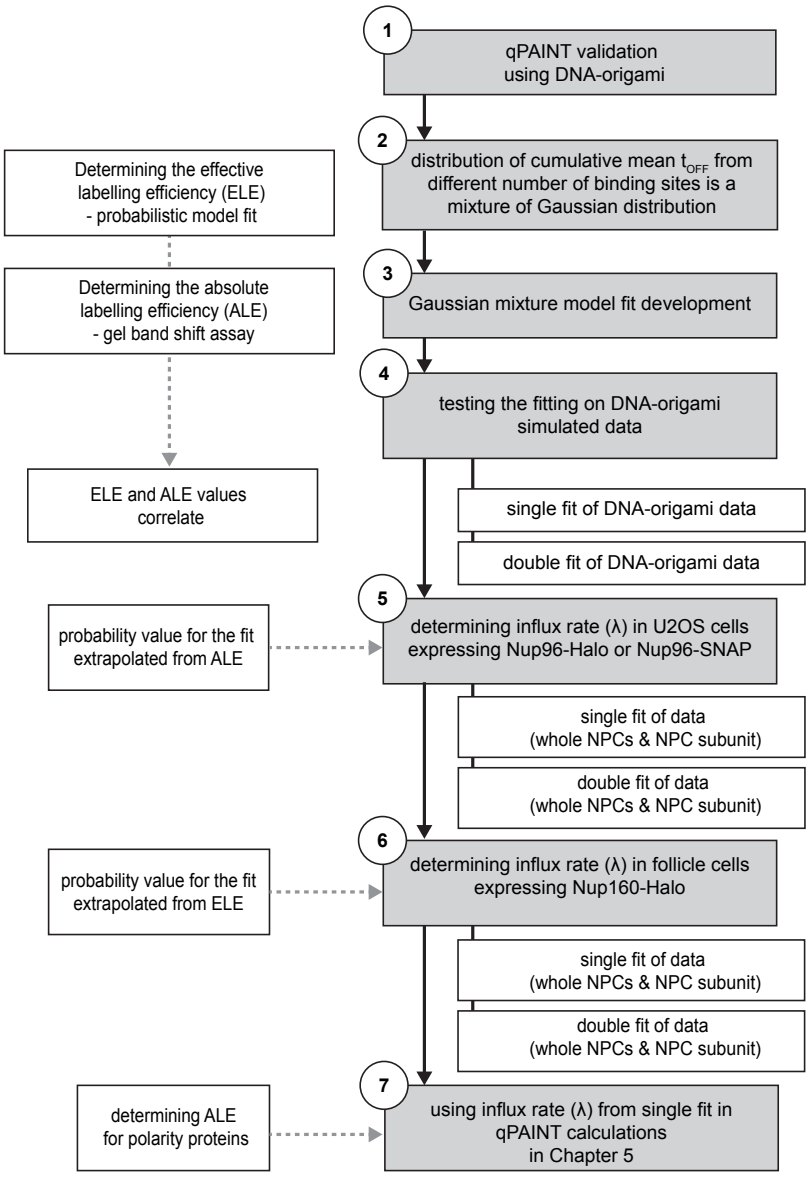


Figure 4.1 **Work-flow of the main experiments presented in this chapter.**

reaction rate upon dissociation. First, the second-order reaction rate for hybridisation means that two reactants are involved, in this case two single-stranded DNA sequences that will hybridise (kinetic constant k_{ON}). Second, the first-order reaction rate (k_{OFF}) means that only one reactant is involved, in this case a complex of two single-strand DNA sequences that dissociates. k_{OFF} is the probability that a DNA hybrid will fall apart per unit time (Pollard 2010).

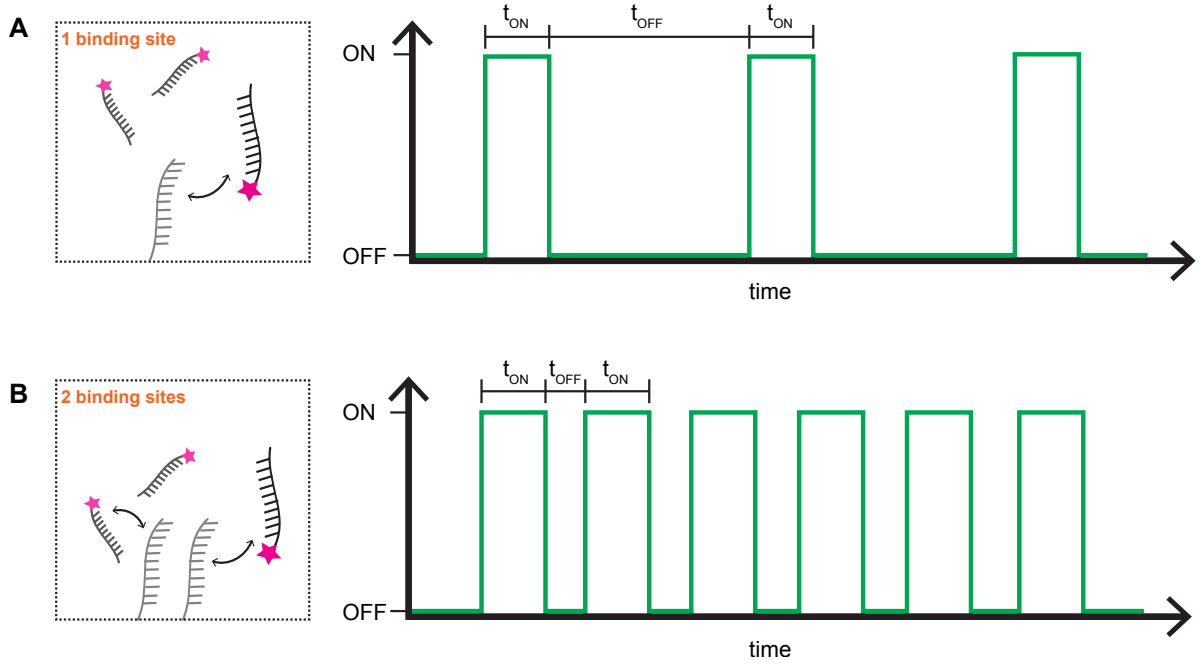


Figure 4.2 **The main principle behind the qPAINT approach – a higher number of the binding sites in an area of interest is reflected by a shorter mean dark time.** (A) Left: a schematic showing an area of interest with a single docking oligo surrounded by freely diffusing imager oligos. Right: a temporal trace of imager oligo binding events for an area containing a single binding site. (B) Left: a schematic showing an area of interest with two docking oligos surrounded by freely diffusing imager oligos. Right: a temporal trace of imager oligo binding events for an area containing two binding sites. While the bright time (t_{ON}) will stay the same, the mean dark time (t_{OFF}) will be shorter.

These kinetic constants determine the length of ON and OFF-times. Because one of the complementary strands is conjugated to a fluorophore it is defined as an imager oligo. The strand that is hybridised by the imager oligo is called the docking oligo (here also referred as a single binding site). The ON time (t_{ON}) is the length of time that the imager oligo is in a complex with the docking oligo, while the OFF time (t_{OFF}) is the length of time between two hybridization events for a respective single binding site. t_{ON} is given by k_{OFF} as $t_{ON} = 1/k_{OFF}$. t_{OFF} is given by the influx rate of imager oligos as $t_{OFF} = 1/\text{influx rate}$, where the influx rate is the number of binding events per single binding site per unit time. More formally influx rate $= k_{ON} \times$ imager concentration. Practically, the influx rate can be calculated by determining the number of binding events per single binding site in a unit of time:

$$\text{influx rate}[s^{-1}] = \frac{\text{number of binding events per single binding site}}{t_{OFF}} \tag{4.1}$$

The rationale behind counting the number of binding sites from the kinetic constants is as following. If a time trace of t_{ON} for a single binding site exhibits a frequency of X and a mean t_{OFF} length of Y, then a time trace for two binding sites exhibits a frequency of 2X and a mean t_{OFF} length of Y/2 (Figure 4.2). Practically, the number of binding sites in the region of interest is calculated by first determining the mean t_{OFF} for this region and then using the following equation:

$$N_{\text{binding sites}} = \frac{1}{\text{influx rate} \times \text{mean } t_{OFF}} \tag{4.2}$$

4.1.2 Experimental design

Validating qPAINT in vitro with DNA origami

As detailed above, qPAINT critically depends on an accurate measure of the influx rate. To better understand how to measure the influx rate and validate our data analysis pipeline I used an artificial structure with known number of binding sites, so-called DNA-origami. DNA-origami has two major advantages as an initial positive control. First, it can be immobilised on cover glass thereby reducing any measurement artefacts due to optical aberrations or sample drift. Second, as a purely synthetic structure it has a well-defined number of binding sites. Having prior knowledge of the number of binding sites is critical in control samples when determining the influx rate. Thus, DNA-origami is an ideal initial positive control as demonstrated before (Jungmann et al. 2016).

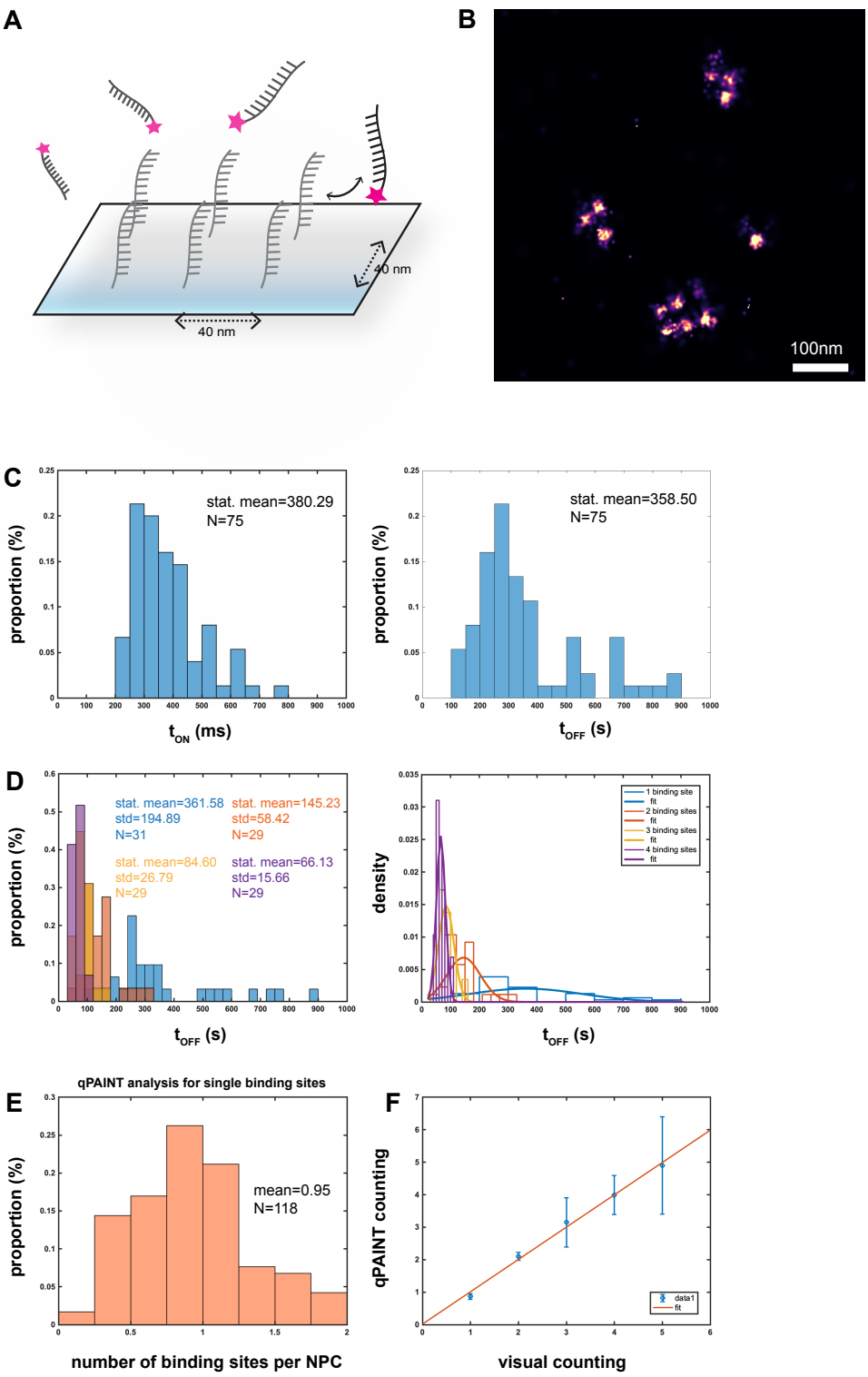


Figure 4.3 **Benchmarking the qPAINT approach on DNA-origami.** Legend on next page.

Figure 4.3 *(previous page)* (A) A schematic showing DNA origami with 6 binding sites, each 40 nm apart, surrounded by freely diffusing imager oligos. (B) An example super-resolution image of 4 different DNA-origami structures with different number of visualised binding sites. (C) Left: a histogram showing the distribution of mean t_{ON} for 75 different single binding sites from DNA origami experiments. Right: a histogram showing the distribution of mean t_{OFF} for 75 different single binding sites from DNA origami experiments. (D) Left: a histogram showing the distribution of mean t_{OFF} when multiple 1, 2, 3 or 4 binding sites were analysed. All four groups are colour-coded. Right: distributions of mean t_{OFF} from the left histogram fitted with the Gaussian distribution. (E) A histogram showing distribution of qPAINT quantification for 118 different single binding sites from DNA origami experiments using the influx rate determined in (C, right). (F) A weighted plot showing a correlation between visually counting the DNA origami structures with different numbers of visualised binding sites and qPAINT quantification result. The linear fit is shown with the red line.

I first imaged DNA origami structures with 6 binding sites with P1 sequence (as shown in Table 2.2 in Chapter 2) that were 40 nm apart (Figure 4.3A). The imager oligo concentration used was 2 nM. The super resolution reconstructions contained DNA origami structures with multiple docking sites visualized (Figure 4.3B). A maximum of five binding sites were visible in the image reconstruction. Lower number of visualised binding sites can happen due to DNA origami damage upon their immobilisation or due to reactive oxygen species produced during imaging (Blumhardt et al. 2018).

To investigate the average binding kinetics for a single binding site, within the DNA-origami experimental system, I first determined mean t_{ON} and t_{OFF} time for 75 different single binding sites using Picasso software. This is a simulation and analysis package that has been made available from Jungmann and co-workers (Schnitzbauer et al. 2017). This software package allows users to simulate DNA paint images and extract mean t_{ON} and mean t_{OFF} in both simulated and experimental data.

Regarding the analysis of the binding kinetics there is an important distinction to be drawn. A mean t_{OFF} for a single binding site is derived from a mean $(1 - 1/e)$ of the exponential function fitted to the cumulative distribution of all t_{OFF} times for aforementioned binding site and will be hereafter referred to as *cumulative mean* t_{OFF}. When I am describing the mean of the distribution of multiple cumulative mean t_{OFF} I am stating a statistical mean, hence it will be hereafter referred to as *statistical mean* t_{OFF}. The mean t_{ON} for a single binding site or for the distribution of multiple mean t_{ON} is always calculated as the *statistical mean* t_{ON} and will be hereafter simply referred to as mean t_{ON}.

The mean t_{ON} time for a single binding site in DNA origami experiments was 380.29 ms (sd=50.0 ms) and the statistical mean t_{OFF} was 358.50 s (sd=78.7 s) resulting in an influx rate of 0.002789 events per binding site per second or 1 binding event every 358.50 seconds (Figure 4.3C).

Investigations of how statistical mean t_{OFF} changes relative to the number of binding sites were also carried out. For a single binding site the statistical mean of the distribution of all cumulative mean t_{OFF} measurements was measured as 361.58 s, compared with 145.23 s for 2 binding sites, 84.60 s for 3 binding sites and 66.13 s for 4 binding sites. Based on these measurements, the distribution of cumulative mean t_{OFF} times appeared to follow Gaussian distribution (Figure 4.3D).

Next, I wondered how accurate would be the qPAINT value for number of binding sites if I used the measured influx rate (358.50 s) and analyse different single binding sites by using their cumulative mean t_{OFF}. Upon analysing 118 different single binding sites the result of qPAINT analysis was on average 0.95 binding sites (sd=0.41) (Figure 4.3E).

Finally, the qPAINT analysis approach was used to calculate the number of binding sites in DNA origami structures having different number of binding sites. The result can be then compared with the visually apparent ground truth. Plotting qPAINT counting result and the visually counted ground truth demonstrated a good agreement between the two with a linear relationship (Figure 4.3F). This confirmed that using qPAINT approach under our experimental settings works as previously reported (Jungmann et al. 2016).

Nuclear pore complex as an in vivo structure for calibration of the influx rate

Since the experimental work in this thesis deals with quantification of the protein number in a tissue sample, one has to determine the influx rate for a single binding site in a particular issue of interest. To use the influx rate from the DNA origami data would likely yield miscalculations for the following two reasons. First, fixed tissue is heavily crosslinked. Heavy crosslinking will slow down the diffusion of the imager oligos relative to a DNA-origami structure where all the binding sites are freely accessible. This makes a comparison between the two systems difficult. Second, because the quality of blinks in an in vitro experiment are very high (high numbers of photon and low background) only a small proportion of blinking events will be rejected during analysis. In a thick tissue, a higher proportion of blinks will be rejected, resulting in longer effective t_{OFF} and thus

lower influx rate. Additionally, the local environment may play a role in the binding times, further complicating the comparison between DNA-origami and in vivo studies.

To calibrate the influx rate under in vivo conditions the nuclear pore complex (NPC) was employed because it exist in the in vivo environment where I planned on counting polarity proteins. Moreover the NPC has a well characterised structure. Here, I used endogenously Halo-tagged Nucleoporin-160 (Nup160) that is present in the cytoplasmic and the nucleoplasmic ring. The stoichiometry of Nup160 is known with 4 protein copies present in each subunit (2 in each ring) yielding 32 copies of Nup160 per NPC (Figure 4.4A) (Weberruss and Antonin 2016).

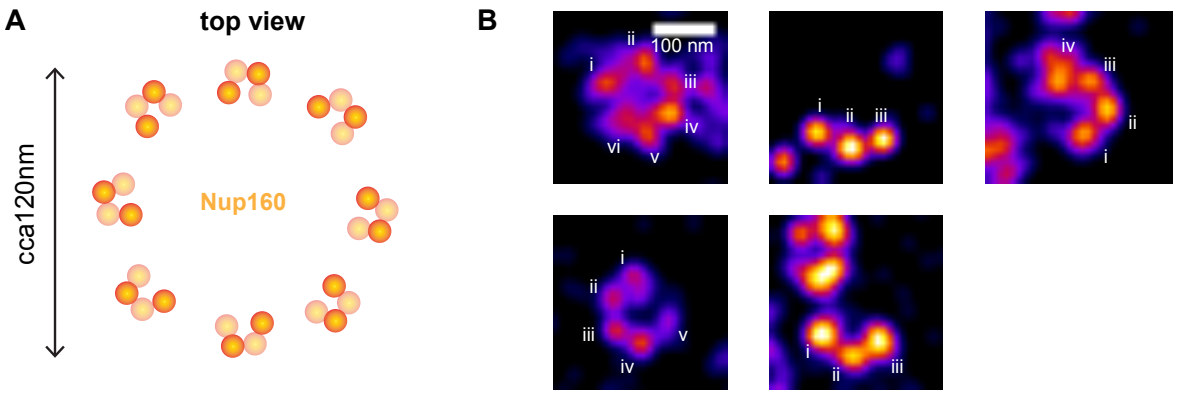


Figure 4.4 **The stoichiometry of Nup160 and its inefficient labelling in vivo.** (A) A schematic showing organisation of Nup160 within the nuclear pore complex viewed from the top. There is 8 subunits and each subunit has 4 copies of Nup160. (B) Five different examples of Nup160-Halo in the follicle cells where single subunits can be distinguished and different labelling efficiency can be observed.

However, even though one can analyse and determine the cumulative mean t_{OFF} for Nup160 in a single NPC subunit, this would still mean 4 copies (or 4 binding sites) of Nup160. In Chapter 3 I showed already that Nup160-Halo is not efficiently labelled (Figure 4.4B). This means that upon analysis of cumulative mean t_{OFF} for different individual NPCs, each NPC has different labelled numbers of Nup160 proteins. Hence the NPC population would exhibit a mixture of different variations of cumulative mean t_{OFF}. In the case of analysis of the whole NPC there are theoretically 32 variations and in the case of analysis of a single NPC subunit there are 4 possible variations. The question is then how can the statistical mean t_{OFF} for a single binding site be determined from a mixture of Gaussian distributions? This I describe in the next section.

Predicting the influx rate for a single binding site from a Gaussian mixture model

Here I present the theory behind the calculation of the statistical mean t_{OFF} (here defined as λ) for a single binding site for which I collaborated with physicist David Jordan. Assume we have a collection of independent Poisson processes each with the same rate parameter λ. The probability of observing n events in a time period t with influx rate constant λ will be given by:

$$P(n = t|\lambda) = \frac{(\lambda t)^n e^{-\lambda t}}{n!} \tag{4.3}$$

Now, in our case, t_{OFF} will be given as Equation (4.3) with n = 0 (no binding events in the interval), or:

$$P(n = t|0) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t} \tag{4.4}$$

In this case we simply have:

$$P(t) = e^{-\lambda t} \tag{4.5}$$

Equation (4.3) is, effectively, the likelihood that we observe a binding event within time interval t. Suppose we measure the t_{OFF} from a single binding site repeatedly and compute the cumulative mean t_{OFF} time. What do we expect that distribution to look like?

The expectation is given by:

$$E(t) = \int_0^\infty tP(t)dt \tag{4.6}$$

but we know from Poisson process that:

$$E(t) = \tau = \frac{1}{\lambda} \tag{4.7}$$

We also know for exponential distribution:

$$\mu = \sigma = \tau = \frac{1}{\lambda} \tag{4.8}$$

By central limit theorem, the distribution of the means of an exponential process should be normally distributed with $\mu = \tau = \frac{1}{\lambda}$ and $\sigma = \frac{\tau}{\sqrt{n}} = \frac{1}{\sqrt{n}\lambda}$, where n is the number of intervals used to compute each exponential rate parameter λ or alternatively the average interval length τ.

For independent exponential distributions the probability of n events is the product of probabilities. So for 2 binding sites the interval time distribution will look like:

$$P(t)^2 = (e^{-\lambda t})^2 = e^{-2\lambda t} \tag{4.9}$$

This will have a Gaussian expected value distribution with $\mu = \tau = \frac{1}{2\lambda}$ and $\sigma = \frac{1}{2\lambda\sqrt{n}}$. Thus if you have $P = (P_1, P_2)$ probability of having 1 or 2 binding sites, the overall distribution will be:

$$P = P(1)\frac{1}{\sqrt{2\pi(\frac{1}{\sqrt{n}\lambda^2}}}e^{-}(\frac{(t-(\frac{1}{\lambda}))^2}{2(\frac{1}{\sqrt{n}\lambda})^2}) + P(2)\frac{1}{\sqrt{2\pi(\frac{1}{\sqrt{n}(2\lambda)^2}}}e^{-}(\frac{(t-(\frac{1}{(2\lambda)}))^2}{2(\frac{1}{\sqrt{n}(2\lambda)})^2}) \tag{4.10}$$

For up to B possible binding sites:

$$P(\tau) = \sum_{i}^{B} p_i \frac{1}{\sqrt{2\pi(\frac{1}{\sqrt{i}\lambda})^2}} exp(-\frac{(\tau-(\frac{1}{i2}))^2}{2(\frac{1}{\sqrt{ni}\lambda})^2}) \tag{4.11}$$

Equation (4.11) matches the overall distribution of the statistical mean t_{OFF}. It is a Gaussian mixture model with B Gaussians, however all B gaussians are determined by a single parameter, λ, and weighted with a parameter p_i (the amplitude of each Gaussian). Here, fitting measured distribution of statistical mean t_{OFF} data with Equation (4.11) where labelling efficiency p_i is also determined will give us the measured value of λ.

Determining the labelling efficiency

Equation (4.11) requests that one need to know the probabilities of respective Gaussian distributions within the mixture. In Equation (4.11) this is the weighting factor p_i. Therefore, it is crucial to determine the labelling efficiency of Nup160 in order to deduce the distribution of binding sites labelled. Assuming that all copies of Nup160 per NPC have the same probability to be labelled, one can use the probabilistic model to determine the labelling efficiency as described before (Thevathasan et al. 2019).

The binomial probability density function:

$$B(k|n,p) = \begin{pmatrix} n \\ k \end{pmatrix} p^k (1-p)^{n-k} \tag{4.12}$$

describes the probability of observing k successes in n independent trials, where the probability of success in any given trial is p. Thus, the probability of a subunit of the NPC (consisting of 4 potentially labelled binding sites) to be dark is $p_{\text{dark}} = B(0|4, p_{\text{label}})$ and the probability to see a corner with at least one site being labelled is $p_{\text{bright}} = 1 - p_{\text{dark}}$. The probability of N out of 8 subunits being visible is:

$$p(N|p_{\text{label}}) = B(N|8, p_{\text{bright}}) = B(N|8, 1 - B(0, 4, p_{\text{label}})) \tag{4.13}$$

Fitting a histogram of the number of visualised subunits of all NPCs with the probabilistic model described in Equation (4.13) can be therefore used to calculate the labelling efficiency. The number of labelled binding sites and subsequently the number of visualized subunits will increase with increasing labelling efficiency (Figure 4.5). This labelling efficiency is here referred as “effective” labelling efficiency (ELE). The ELE is defined as labelling efficiency

obtained from visualised NPC subunits. They were visualised because of hybridisation events with the imager oligos.

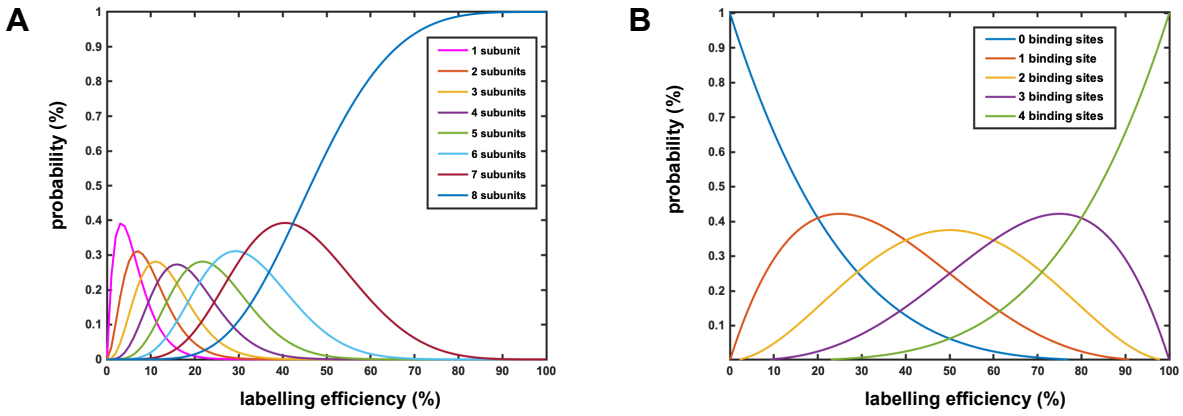


Figure 4.5 **The labelling efficiency affects the number of labelled subunits per NPC and labelled binding sites per NPC subunit.** (A) A plot showing how the number of labelled subunits per NPC changes with increasing labelling efficiency. (B) A plot showing how the number of labelled binding sites per NPC subunit changes with increasing labelling efficiency.

This ELE can be used to calculate the expected probabilities of binding sites labelled per NPC or per subunit, respectively. For example, with 50% ELE around 14% of NPCs would have 16 binding sites labelled, while around 40% of the subunits would have 2 binding sites labelled (Figure 4.6). It is necessary to calculate these probabilities (p_i) to calculate the influx rate (λ) as described in Equation (4.11).

Validating the Gaussian mixture model on simulated data

To validate the model described in Section 4.1.2, simulated super-resolution images of DNA origami were generated in Picasso software (Schnitzbauer et al. 2017). For simulated data the cumulative mean t_{OFF} for a single binding site was set arbitrary to 500 s with 4 nM concentration of the imager strand and 250 ms exposure time (4 frames per second).

Analysis of the simulated data showed that the statistical mean t_{OFF} for a single binding site was, on average, 571.95 s, with median value (509.97 s) being much closer to the simulated mean t_{OFF} (Figure 4.7A). Plotting distributions of the cumulative mean t_{OFF} for 1, 2 and 3 binding sites revealed that cumulative mean t_{OFF} times follow Gaussian distribution as shown before on much smaller sample from experimental DNA origami data (Figure 4.7A).

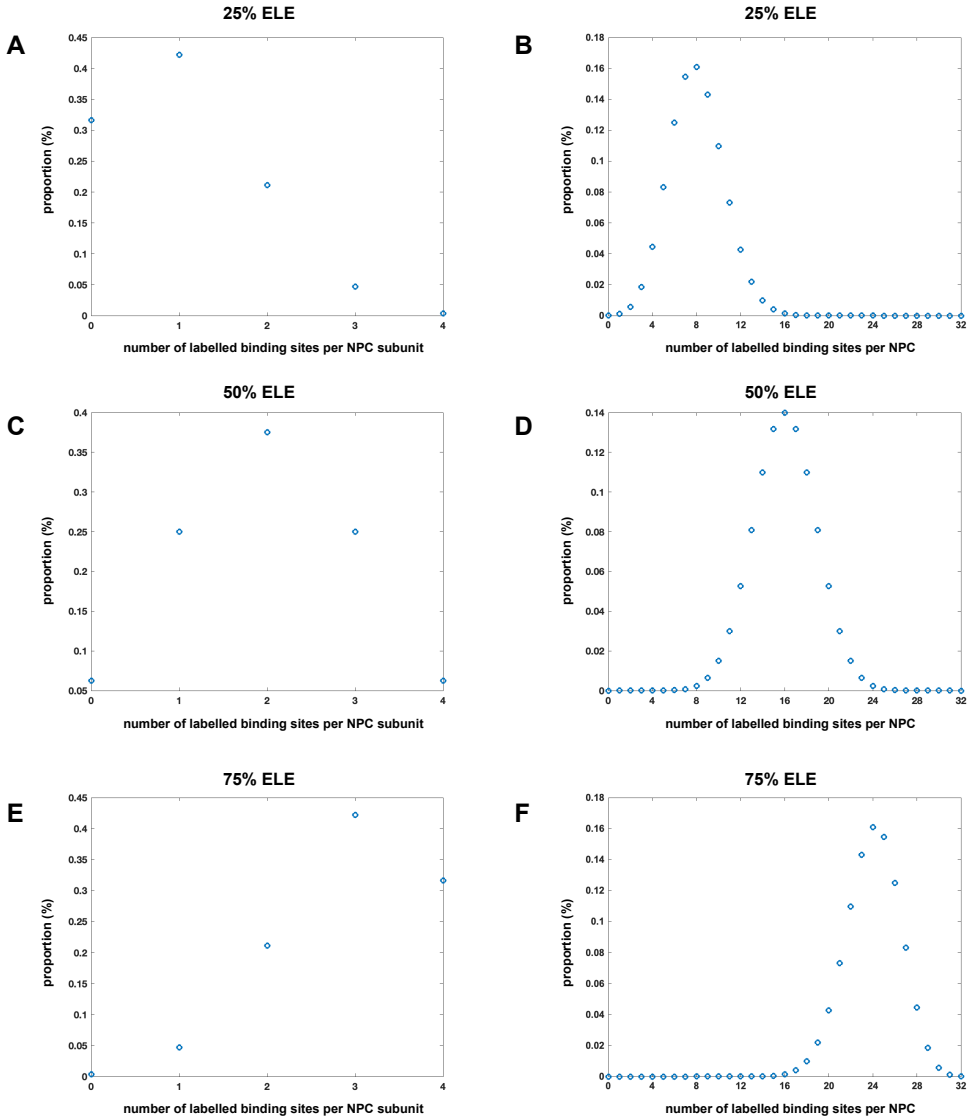


Figure 4.6 **An example of how effective labelling efficiency influences the number of labelled binding sites per NPC and its subunit.** (A) A plot showing the distribution of the number of labelled binding sites per NPC subunit with 25% effective labelling efficiency. (B) A plot showing the distribution of the number of labelled binding sites per NPC with 25% effective labelling efficiency. (C) A plot showing the distribution of the number of labelled binding sites per NPC subunit with 50% effective labelling efficiency. (D) A plot showing the distribution of the number of labelled binding sites per NPC with 50% effective labelling efficiency. (E) A plot showing the distribution of the number of labelled binding sites per NPC subunit with 75% effective labelling efficiency. (F) A plot showing the distribution of the number of labelled binding sites per NPC with 75% effective labelling efficiency.

Despite a seemingly large sample size, the statistical mean t_{OFF} was not in good agreement with the ground truth value I calculated based on statistical mean and median values in three different samples sizes in the simulated data (Figure 4.7B). Again I observed that

median value is much closer to the simulated cumulative mean t_{OFF}, however a factor of ten increase in the sample size did not drastically shift the mean or the median value towards the ground truth value. This is likely due to the presence of outliers in the dataset, hence also the reason why the median value is closer to the ground truth.

Next I performed fitting of cumulative mean t_{OFF} distribution from simulated DNA origami data using Equation (4.11) to determine λ parameter (the influx rate). The fitting was performed using Matlab (see Appendix E). Because Equation (4.11) has only a single parameter, the influx rate, that is allowed to vary I refer to this approach as "single" fitting. The cumulative mean t_{OFF} distribution of simulated DNA origami data had a mixture of cumulative mean t_{OFF} from a single binding site (75% of the dataset points), 2 binding sites (24% of the dataset points) and 3 binding sites (1% of the dataset points). These mixture corresponded to 20% labelling efficiency upon 4 possible binding sites (Figure 4.5B). 20% labelling efficiency was chosen arbitrarily. I tested the fitting on two different sample sizes. The calculated λ was 538 s for the sample size of 100 simulated DNA origami structures. Increasing the number of simulated DNA origami structures shifted the measured λ closer to 515 s (Figure 4.7C and Figure 4.7D).

To see if fitting would perform better with probabilities (from which the labelling efficiency can be derived) as an additional parameter the simulated data was fitted with Equation (4.11) allowing both the influx rate (λ) and labelling efficiency (p_1) to vary. Because in this case two parameters, λ and labelling efficiency (p_i), were allowed to vary I refer to this approach as "double" fitting (see Appendix F). Here, the measured λ was 471 s and predicted labelling efficiency was 12%. Hereafter, the *predicted labelling efficiency* is independent of the initial measurement of the experimental labelling efficiency and is only used as a consistency check.

Continuing with the double fit on the bigger simulated sample size yielded λ of 520 s and predicted labelling efficiency of 23%. This results suggested that a single fit is superior when calculating λ value in datasets having either small or big sample size.

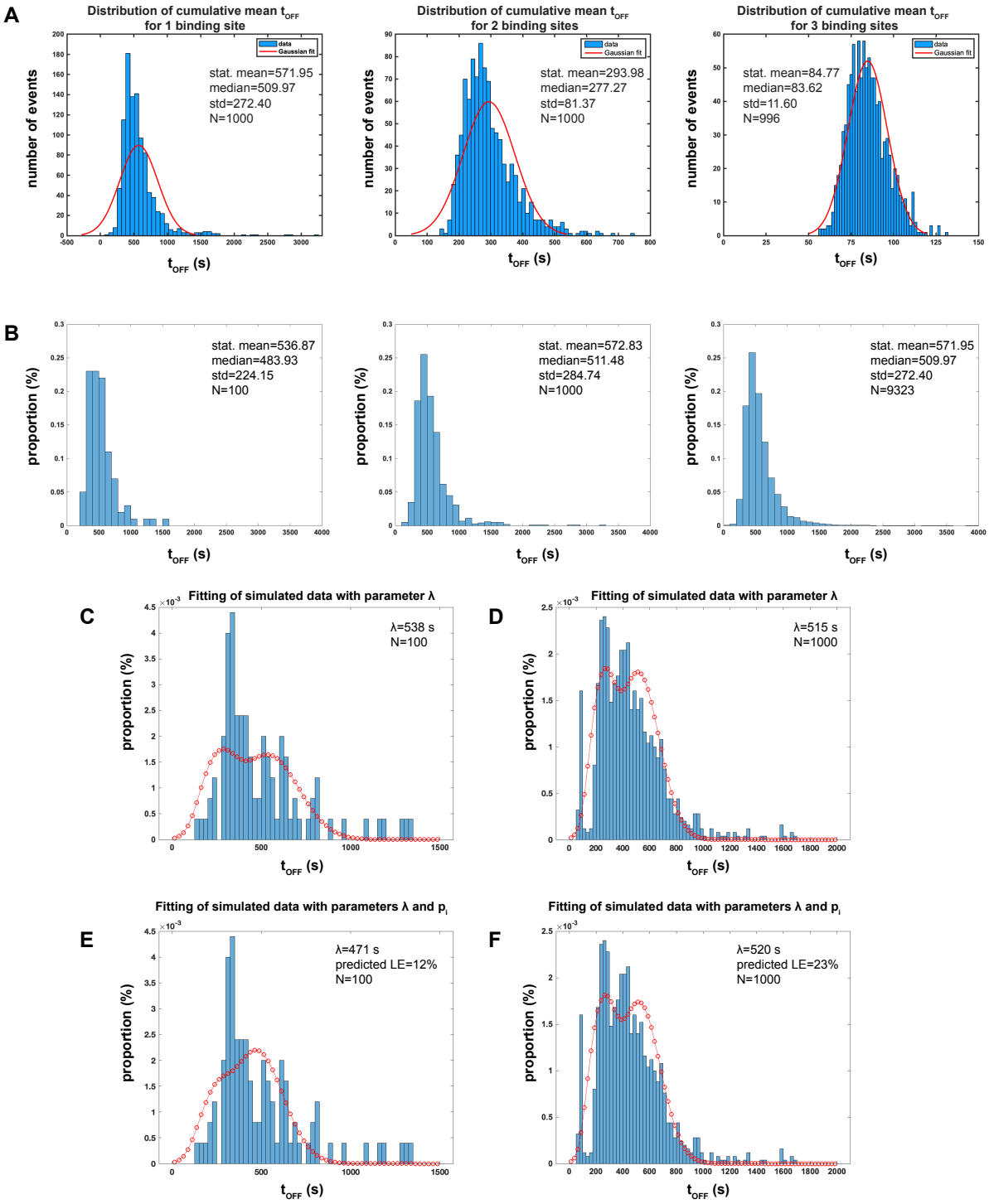


Figure 4.7 **Determining the influx rate for a single binding site from the simulated DNA origami experiments.** Legend on next page.

Figure 4.7 *(previous page)* (A) A histogram showing the distribution of cumulative mean t_{OFF} for an area containing 1 (left), 2 (middle), or 3 (right) binding sites, respectively, from the simulated DNA origami experiments. The means stated in the plots are statistical means. The distributions were fitted with the Gaussian fit (the red line). (B) Histograms showing the distribution of cumulative mean t_{OFF} for 100 (left), 1000 (middle), or 9323 (right) single binding sites from the simulated DNA origami experiments. The means stated in the plots are statistical means. (C and D) Fitting (the red line) with a single parameter λ of the distributions of cumulative mean t_{OFF} for 100 different single binding sites (C) and 1000 different single binding sites (D) from the simulated DNA origami experiments. (E and F) Fitting (the red line) with parameters λ and p_i of the distributions of cumulative mean t_{OFF} for 100 different single binding sites (E) and 1000 different single binding sites (F) from the simulated DNA origami experiments.

4.2 Results

4.2.1 Correlating the effective and the absolute labelling efficiency of investigated proteins

To perform fitting of mean t_{OFF} distribution for Nup160 data I first had to determine its labelling efficiency since labelling efficiency is not accurately determined from the fitting when used as a fit parameter as showed in Section 4.1.2.

The effective labelling efficiency (ELE) can be easily determined for the NPCs since their 8 subunits are radially arranged with a known number of binding sites per subunit. From counting the number of visualised subunits per NPC in the super-resolution images, the ELE can be deduced from a statistical analysis with a probabilistic model as described in Section 4.1.2. However, my ultimate goal is to count polarity proteins, which do not form any structures of a known stoichiometry, and therefore the same approach cannot be used to determine their labelling efficiency. Using a Western blot-based gel band shift assay (the method is explained in Chapter 2) one can quantify the absolute labelling efficiency (ALE).

Using a Western blot-based gel band shift assay one can quantify the absolute labelling efficiency (ALE). This method utilises the fact that tagged protein that was labelled has approximately 3 kDa bigger molecular weight (due to ligand conjugated to the docking oligo) than non-tagged protein. The gel band shift assay works first by running the labelled sample on a gel and then preforming a Western blot for Halo- or SNAP-tag. The band containing the labelled protein fraction will be shifted from the unlabelled protein fractions

(Figure 4.8) The Western blot is a fluorescent Western blot where the signal intensity is linearly correlated with the protein abundance. This enables band's fluorescence signal quantification and calculation of the signal ratio between the two bands to asses ALE.

The ALE differs from ELE by stating the absolute proportion of protein molecules that were labelled with the docking oligo. On the other hand, the ELE is stating the proportion of protein molecules that were hybridised with the imager oligo and hence visualised.

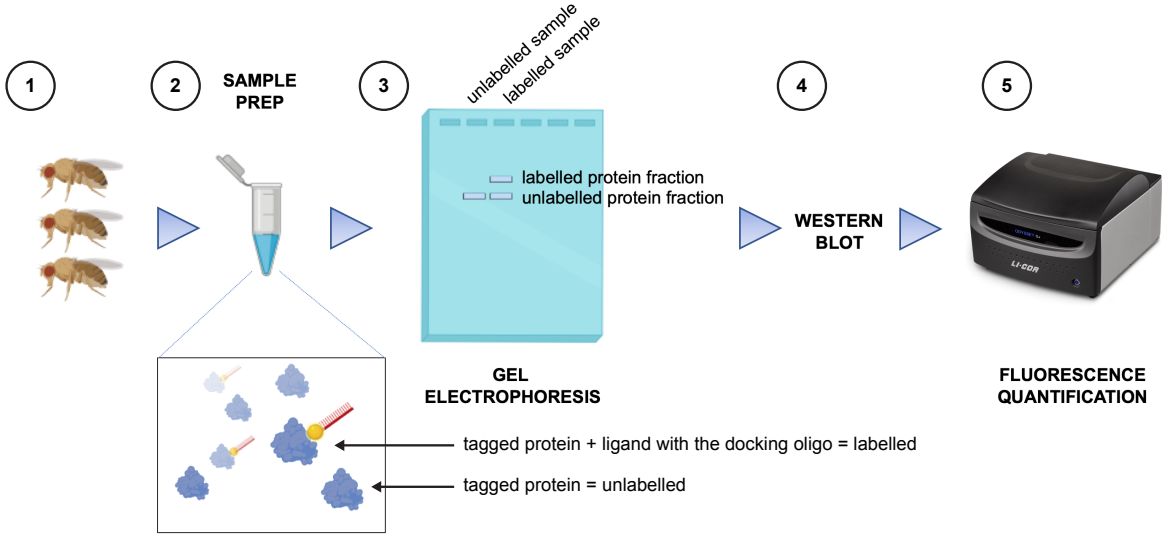


Figure 4.8 **A simplified schematic of the gel band shift assay work-flow.** 1: Fly ovaries are isolated from flies that endogenously express tagged (SNAP or Halo) protein of interest. 2: The ovaries are labelled with respective ligand in a tube and a protein lysate is prepared. 3: A protein sample is run on a gel electrophoresis, the labelled protein fraction is heavier and travels through the gel at the slower speed. 4: Fluorescent western blot is performed. 5: The fluorescent signal intensity from the bands are machine-detected and quantified.

I first wanted to investigate how the ALE relates to the ELE. If the ELE closely matches the ALE, the gel band shift assay can be used for any protein to asses labelling efficiency. The ALE can be later used to quantify the super-resolution images of the investigated protein molecules and calculate the absolute number of proteins molecules in a region of interest.

To this end, I first used U2OS cells in which Nup96 was endogenously-tagged with either SNAP or Halo tag, respectively (Thevathasan et al. 2019). The cells were fixed and labelled with the respective ligand conjugated to the docking oligo. Next, the cells were

imaged using the imager oligo conjugated to Cy3B with TIRF imaging, which achieves better super resolution than widefield imaging since only a thin section of the specimen is illuminated. The NPCs were labelled incompletely, which was a sign of a low labelling efficiency (Figure 4.9A).

To systematically approach the counting of NPC subunits, an automated image analysis tool was developed, following a similar approach to that previously described (Thevathasan et al. 2019). The NPCs were automatically identified and then divided into 8 segments (representing the 8 subunits of the NPC).

The position of segments was done by calculating the angular offset giving the maximum detection frequency in any of eight segment bins. The final segment divisions are placed at $s\frac{\pi}{4} + offset(maxf) - \frac{\pi}{8}$ for $s = 0.7$, giving eight segments of equal size where one contains the maximum possible number of detections. The number of segments containing at least 10 localisations was then used to plot a distribution of visualised subunits and fit the probabilistic model as showed in Section 4.1.2 (Figure 4.9B).

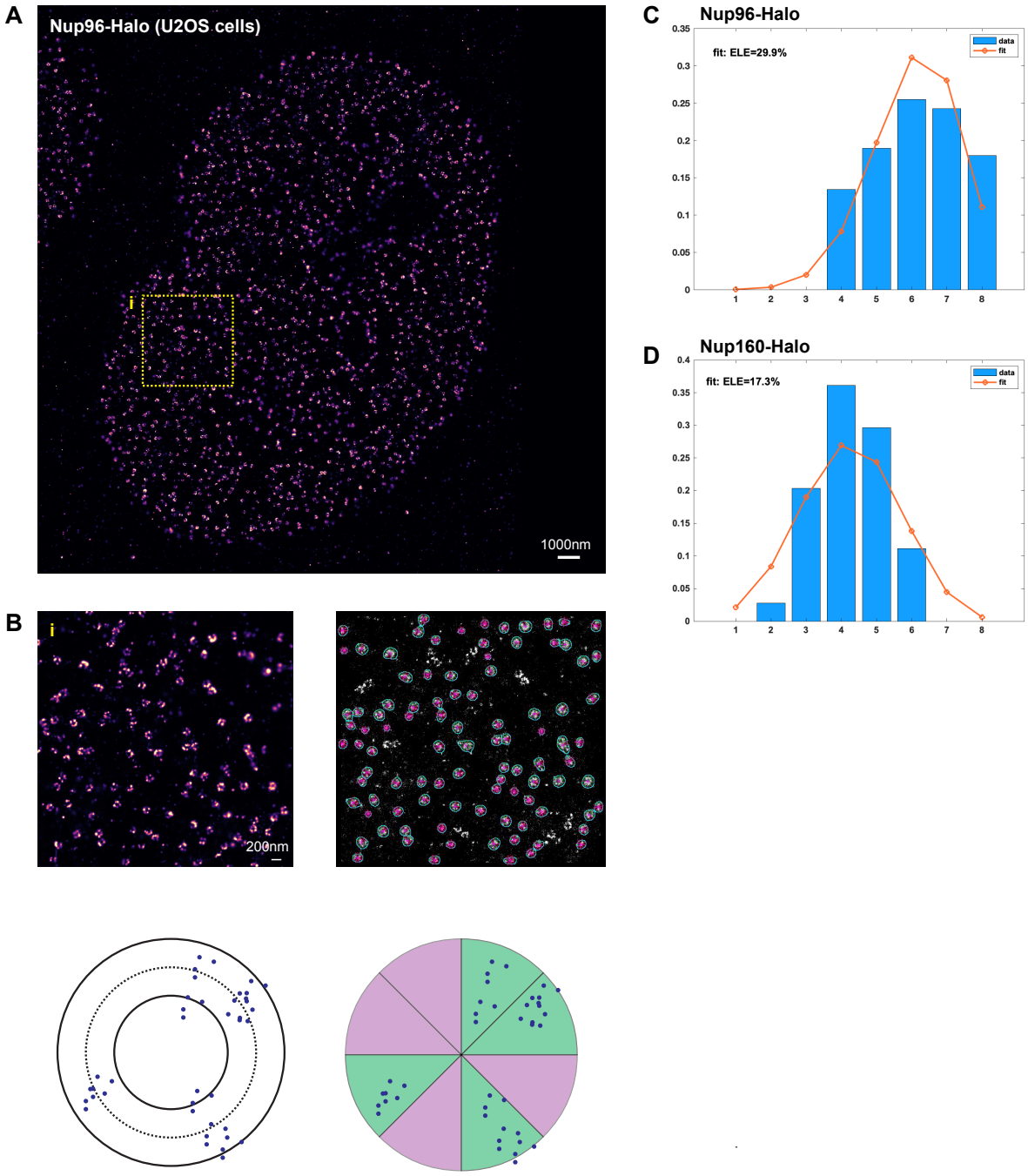


Figure 4.9 **Determining the effective labelling efficiency of nucleoporin proteins in U2OS cultured cells and fruit fly follicular cells.** Legend on next page.

Figure 4.9 *(previous page)* (A) An example of a super-resolution image of Nup96-Halo in a cultured U2OS cell. The image is a reconstruction of 25000 frames. The imager oligo concentration used was 2 nM. The yellow-lined box outlines the region shown at higher magnification in (B). (B) Top left: The zoomed-in part of the image from (A). Top right: The automatic segmentation of single NPCs and the analysis of their labelling efficiencies. Bottom left: A circle is fitted to each segmented NPC and all localisations between the inner ring (2r=60 nm) and the outer ring (2r=140 nm) are used for the analysis. Bottom right: The segmented NPCs are sliced into 8 subsegments and the number of segments containing at least 10 localisations are counted. (C) An example of a histogram showign the number of visualised NPC subunits for a Nup96-Halo sample fitted with a probabilistic model to obtain the effective labelling efficiency. (D) An example of a histogram showing the number of visualised NPC subunits for a Nup160-Halo sample fitted with a probabilistic model to obtain the effective labelling efficiency.

Using reconstructed super resolution images of Nup96-Halo, the ELE was determined to be 29.9% (sd=7.1) (Figure 4.9C). The ALE was measured as 32% (sd=7.8) using the gel band shift assay as detailed above (Figure 4.10). This numbers are in reasonable agreement when considering their respective uncertainties. For Nup96-SNAP the ELE was 21% (sd=3.17) (Figure 4.10D). For Nup96-SNAP we were unable to measure the ALE as the gel band shift of the labelled fraction appeared to blurry and was thus not separated from the unlabelled fraction (Figure 4.10B). Nevertheless, the comparison between ELE and ALE for Nup96-Halo suggests that they are similar (Figure 4.10D). Thus, a gel band shift assay can be used to quantify the labelling efficiency of the protein of interest.

Next, I wanted to repeat the same analysis in the fruit fly follicular tissue endogenously expressing Nup160-Halo. Because the nuclei are much deeper TIRF imaging could not be used. Unfortunately, the automated image analysis could not be done since the NPC are clustered in this cell type, moreover the level of unspecific signal was much higher than in cultured cells, making the segmentation difficult. Therefore, I manually analysed NPCs and counted the subunits. The ELE was determined to be 17.2% (n=107) (Figure 4.9D).

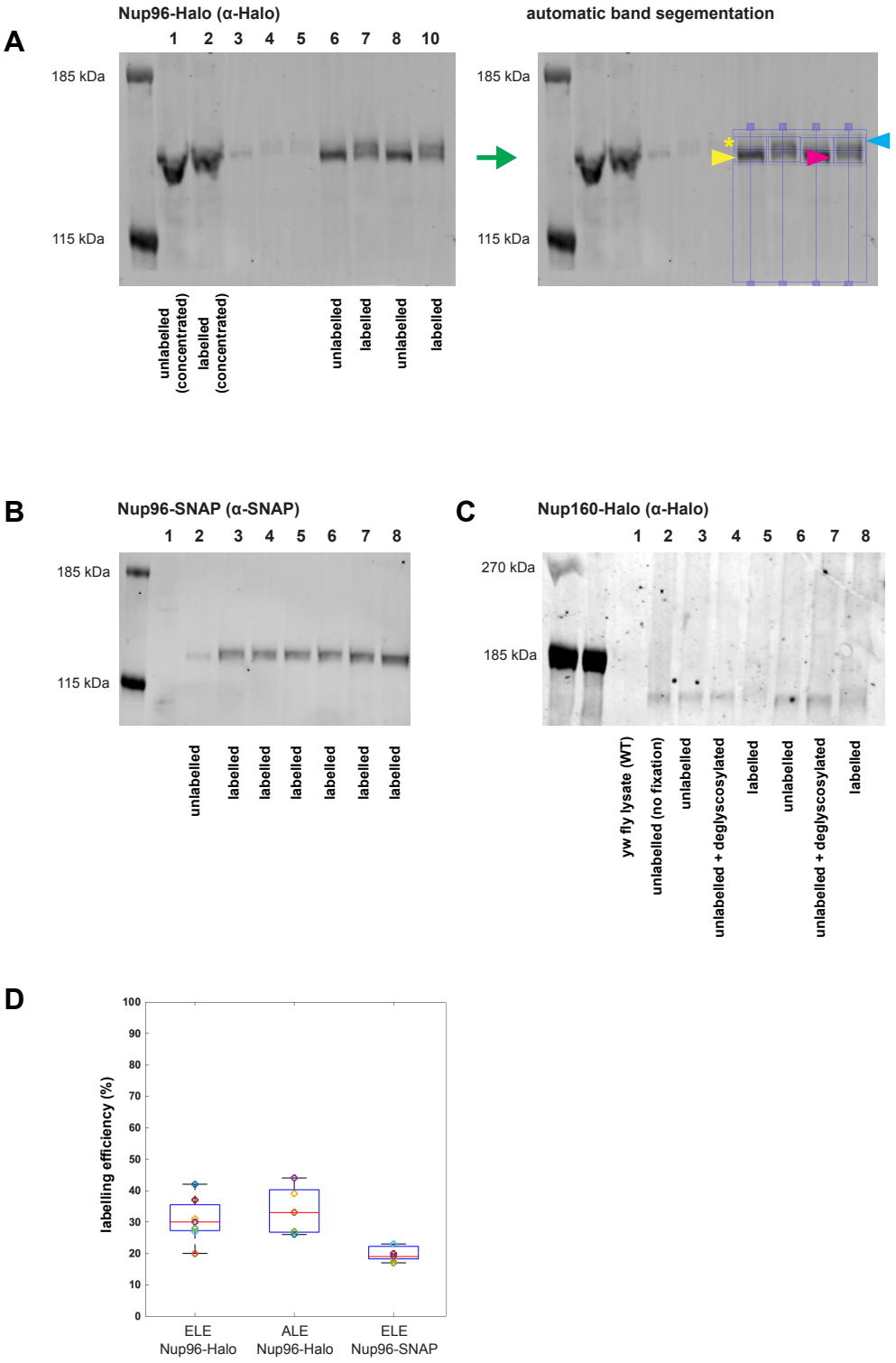


Figure 4.10 **Determining the absolute labelling efficiency for nucleoporin proteins.** Legend on next page.

Figure 4.10 *(previous page)* (A) Left: A Western blot of a lysate from fixed U2OS cells expressing Nup96-Halo using an antibody against the Halo epitope. Lane 1 (far left): a concentrated protein lysate from the unlabelled cells. Lane 2: a concentrated protein lysate from labelled cells using Halo ligand conjugated to the P1 docking oligo. Lane 3-5: blank. Lane 6: a protein lysate from unlabelled cells. Lane 7: a protein lysate from labelled cells. Lane 8: a protein lysate (higher protein concentration) from unlabelled cells. Lane 9: a protein lysate (higher protein concentration) from labelled cells. Right: Automatic band segmentation of a typical Western blot. The yellow arrow indicates a band of unlabelled Nup96-Halo protein and the yellow asterisk marks the background from the band bellow. The magenta arrow indicates the band of an unlabelled fraction of Nup96-Halo protein. The blue arrow marks the band of a labelled fraction of Nup96-Halo protein (the band is shifted because the labelled fraction is approximately 3kDa heavier due to the additional molecular weight of the conjugated docking oligo). (B) A Western blot of a lysate from fixed U2OS cells expressing Nup96-SNAP using an antibody against the SNAP epitope. Lane 1: blank. Lane 2: a protein lysate from unlabelled cells. Lane 3-8: a protein lysate from labelled cells (with increasing protein concentration). Note that the band shift is present but the two bands cannot be distinguished. (C) A Western blot of a lysate from fixed fruit fly follicle cells expressing Nup160-Halo using an antibody against the Halo epitope. Lane 1: a protein lysate from yellow-white flies (wild type, not expressing Halo-tagged Nup160). Lane 2: a protein lysate from unfixed and unlabelled follicle cells. Lane 3: a protein lysate from fixed and unlabelled follicle cells. Lane 4: a deglycosylated protein lysate from unlabelled follicle cells. Lane 5: a protein lysate from labelled follicle cells. Lane 6: a protein lysate (higher concentration) from unlabelled follicle cells. Lane 7: a deglycosylated protein lysate (higher concentration) from unlabelled follicle cells. Lane 8: a protein lysate (higher concentration) from labelled follicle cells. Note that the band shift is present but the two bands cannot be distinguished. (D) A plot showing the absolute and effective labelling efficiencies for Nup96-Halo and Nup96-SNAP samples. Note that there are no data for absolute labelling efficiency for Nup96-SNAP since the band shift could not be distinguished and quantified (see (B)).

Next, the ALE of Nup160-Halo was assed using the Western blot. However, because of the size of the tagged Nup160 (cca 190kDa), the 3kDa band shift was not able to be distinguished. Deglycosylation treatment of the protein lysate did not improve resolution (Figure 4.10C). The thrombin enzyme treatment of the protein lysate was undertaken. This would cut the Nup160 at multiple sites but leave the Halo tag intact. However, unspecific cutting of the Halo tag was observed (data not shown).

Taken together, assessment of the ELE and ALE of the Nup96-Halo provided an evidence that they correlate. This suggests that the gel band shift assay can be used to quantify the absolute labelling efficiency.

4.2.2 Calculating the influx rate for a single binding site in cultured cells

With the information about the labelling efficiency, I could next move to determine the influx rate for a single binding site in U2OS cultured cells endogenously expressing Nup96-Halo or Nup96-SNAP. All experimental results described in the next paragraphs are summarised in Table 4.1, which might help facilitate reading of this chapter.

Using Picasso software I first automatically segmented NPCs and assessed the cumulative mean t_{OFF} of each single NPC (n=1278). Next, the distribution of the cumulative mean t_{OFF} times was plotted and fit using a single parameter (λ), as described above and presented in (Figure 4.11A). The expected probabilities of the number of labelled binding sites were calculated from the previously determined ALE, which was 32%. λ was calculated as 414 s, which translated to an influx rate of 0.002415 s^{-1}. This influx rate was used to calculate the number of binding sites in previously segmented NPC images using Picasso software. The average number of binding sites per NPC was 10.34, which would translate into 32% of predicted labelling efficiency (Figure 4.11B). The predicted labelling efficiency based on the qPAINT analysis matched well with the ALE assayed in the gel band shift assay.

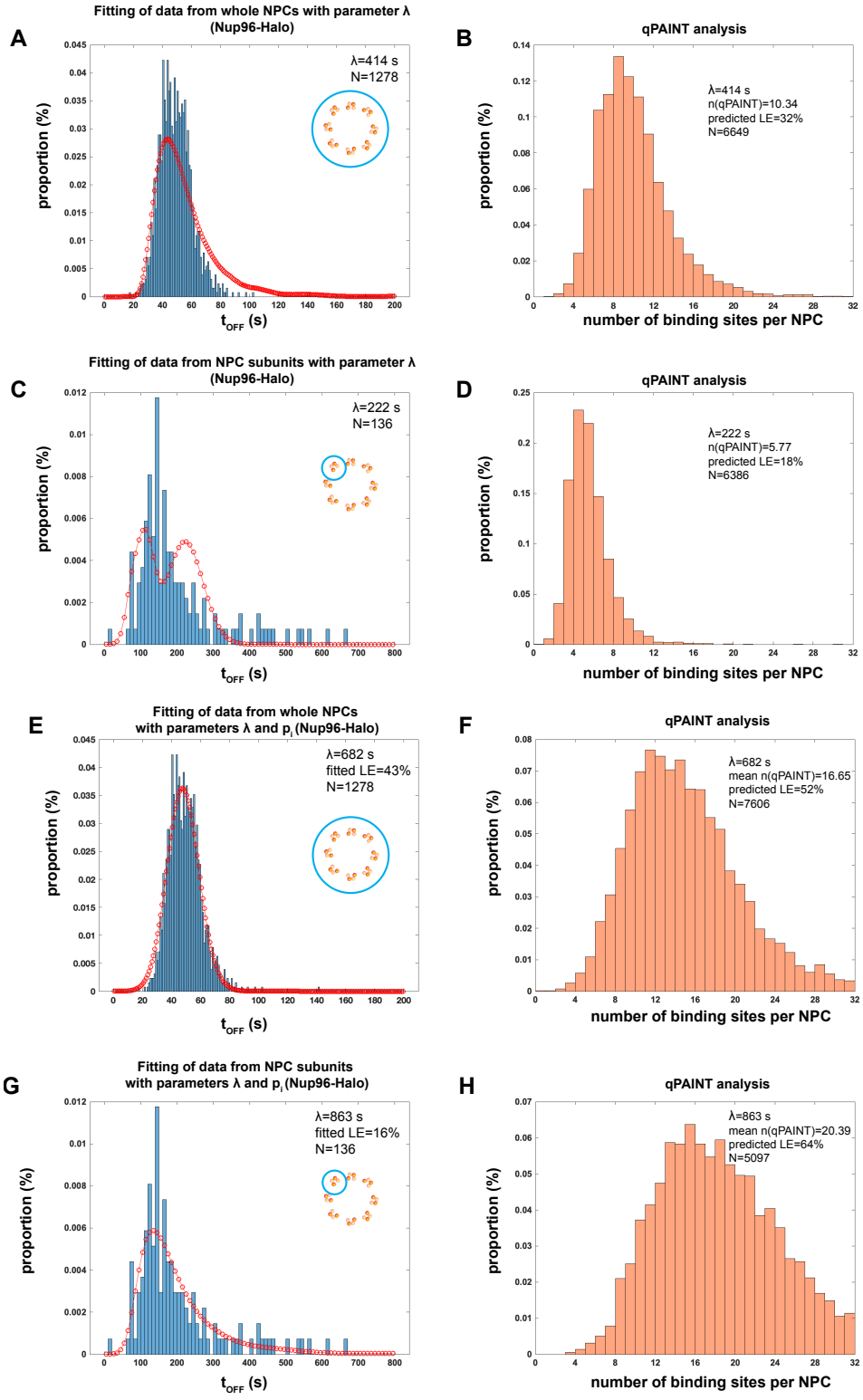


Figure 4.11 **Determining the influx rate for a single binding site by fitting Nup96-Halo cumulative mean t_{OFF} obtained in U2OS cells.** Legend on next page.

Figure 4.11 *(previous page)* (A) Fitting (the red line) with a single parameter λ of the distribution of cumulative mean t_{OFF} for data of whole NPCs. The probability (p_i) of the number of labelled binding sites was calculated from the effective labelling efficiency (32%). (B) The distribution of the number of binding sites per NPC as calculated with qPAINT using the influx rate determined in (A). (C) Fitting (the red line) with a single parameter λ of the distribution of cumulative mean t_{OFF} for data of NPC subunits. The probability (p_i) of the number of labelled binding sites was calculated from the effective labelling efficiency (32%).(D) The distribution of the number of binding sites per NPC as calculated with qPAINT using the influx rate determined in (C). (E) Fitting (the red line) with parameters λ and (p_i) of the distribution of cumulative mean t_{OFF} for data of whole NPCs. The fitted labelling efficiency (LE) was calculated as 43%. (F) The distribution of the number of binding sites per NPC as calculated with qPAINT using the influx rate determined in (E). (G) Fitting (the red line) with parameters λ and (p_i) of the distribution of cumulative mean t_{OFF} for data of NPC subunits. The fitted labelling efficiency (LE) was calculated as 16%. (H) The distribution of the number of binding sites per NPC as calculated with qPAINT using the influx rate determined in (G).

The same analysis was performed on cumulative mean t_{OFF} data from NPC subunits (n=136, manually chosen). λ was calculated as 222 s, which translated to an influx rate of 0.004504 s^{-1} (Figure 4.11C). The average number of binding sites per NPC was measured as 5.77, which would translate into 18% of predicted labelling efficiency (Figure 4.11D).

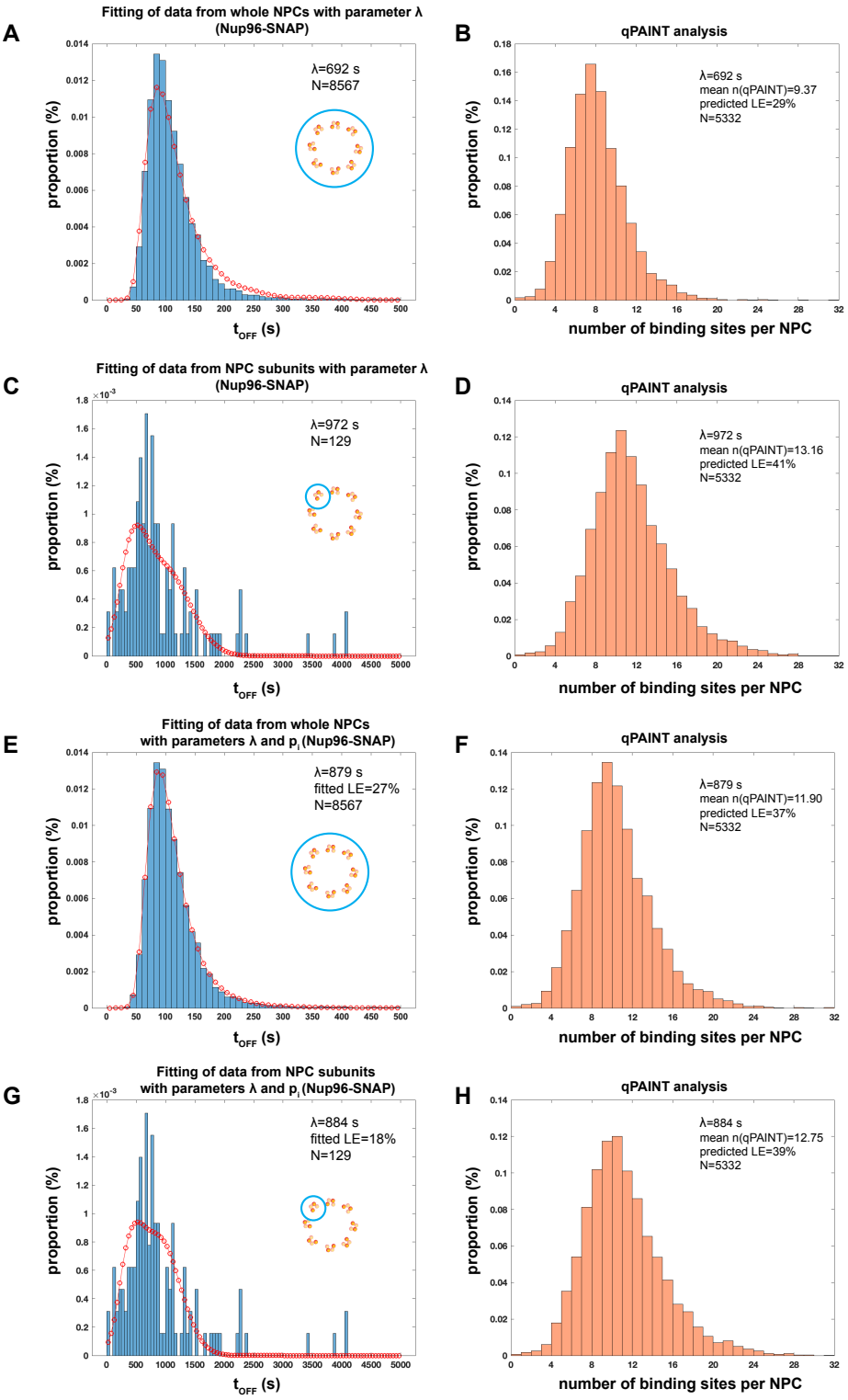


Figure 4.12 **Determining the influx rate for a single binding site by fitting of Nup96-SNAP cumulative mean t_{OFF} obtained in U2OS cells.** Legend on the next page.

Figure 4.12 *(previous page)* (A) Fitting (the red line) with a single parameter λ of the distribution of cumulative mean t_{OFF} for data of whole NPCs. The probability (p_i) of the number of labelled binding sites was calculated from the effective labelling efficiency (20%). (B) The distribution of the number of binding sites per NPC as calculated with qPAINT using the influx rate determined in (A). (C) Fitting (the red line) with a single parameter λ of the distribution of cumulative mean t_{OFF} for data of NPC subunits. The probability (p_i) of the number of labelled binding sites was calculated from the effective labelling efficiency (20%).(D) The distribution of the number of binding sites per NPC as calculated with qPAINT using the influx rate determined in (C). (E) Fitting (the red line) with parameters λ and (p_i) of the distribution of cumulative mean t_{OFF} for data of whole NPCs. The fitted labelling efficiency (LE) was calculated as 27%. (F) The distribution of the number of binding sites per NPC as calculated with qPAINT using the influx rate determined in (E). (G) Fitting (the red line) with parameters λ and (p_i) of the distribution of cumulative mean t_{OFF} for data of NPC subunits. The fitted labelling efficiency (LE) was calculated as 18%. (H) The distribution of the number of binding sites per NPC as calculated with qPAINT using the influx rate determined in (G).

To assess how λ would change if the distribution of cumulative mean t_{OFF} were fitted with the two parameter fit, along with λ also with the best probabilities for the number of binding sites (from which a corresponding labelling efficiency can be derived) as a second parameter. For the data from the whole NPCs, the best fit corresponded to parameter λ of 682 s (influx rate=0.001466 s^{-1}) and 43% fitted labelling efficiency (Figure 4.11E). The average number of binding sites per NPC calculated with this influx rate was 16.65, which would translate into 52% predicted labelling efficiency (Figure 4.11F). Here the "fitted labelling efficiency" refers to the value obtained from the fit, while the "predicted labelling efficiency" refers to the value obtained from the qPAINT analysis of the number of binding sites. For the data from the NPC subunits the best fit corresponded to parameter λ of 863 s (influx rate=0.001158 s^{-1}) and 16% fitted labelling efficiency (Figure 4.11G). The average number of binding sites per NPC calculated with this influx rate was 20.39, which would translate into 64% predicted labelling efficiency (Figure 4.11H).

The same analysis was performed on the data from Nup96-SNAP. First using fitting with a single free parameter λ and then using fitting with two free parameters λ and p_i, as described below (Figure 4.12).

For the data from the entire NPCs, λ was measured as 692 s, which translated into influx rate of 0.001445 s^{-1} (Figure 4.12A). The average number of binding sites per NPC using this influx rate was 9.37, which would translate into 29% predicted labelling efficiency (Figure 4.12B). For the data from the NPC subunits the λ was meausred as 972 s or

an influx rate of 0.001028 s^{-1} (Figure 4.12C). The average number of binding sites per NPC using this influx rate was 13.16, corresponding to a 41% predicted labelling efficiency (Figure 4.12D).

Subsequently, λ was calculated using a two parameter fit (Equation (4.11) , allowing p_i and λ to vary). For the data from the entire NPCs, a two parameter fit resulted in a λ of 879 s (influx rate=0.001137 s^{-1}) and 27% fitted labelling efficiency (Figure 4.12E). The average number of binding sites per NPC calculated with this influx rate was 11.90, corresponding to a 37% predicted labelling efficiency (Figure 4.12F). For the data from the NPC subunits the double fit yielded a similar λ of 884 s (influx rate=0.001131 s^{-1}) and 18% fitted labelling efficiency (Figure 4.12G). The average number of binding sites per NPC calculated with this influx rate was 12.75, which would translate into 39% predicted labelling efficiency (Figure 4.12H).

All together these results suggest that the fitting with a single parameter on the data from the whole NPCs is the best way to most accurately calculate the parameter λ as suggested by simulated data as well.

4.2.3 Calculating the influx rate for a single binding site in the fruit fly tissue sample

After establishing an analysis workflow to determine the influx rate for a single binding site in cultured cells, I wanted to repeat the same analysis on the NPCs in the fruit fly egg chambers expressing Nup160-Halo. Here the expected probabilities of the number of labelled binding sites were calculated from the previously determined ELE (17.2%) (Figure 4.9D), since ALE could not be assayed.

NPCs were automatically segmented and the cumulative mean t_{OFF} for each individual NPC was measured (n=3276). Then I plotted the distribution of all cumulative mean t_{OFF} and used single fitting described above to determine the parameter λ. λ was calculated as 464 s, which translated into influx rate of 0.002155 s^{-1} (Figure 4.13A). This influx rate was then used to calculate the number of binding sites in previously segmented NPCs using Picasso software. The average number of binding sites per NPC was 7.63, which would translate into 24% of predicted labelling efficiency (Figure 4.13B).

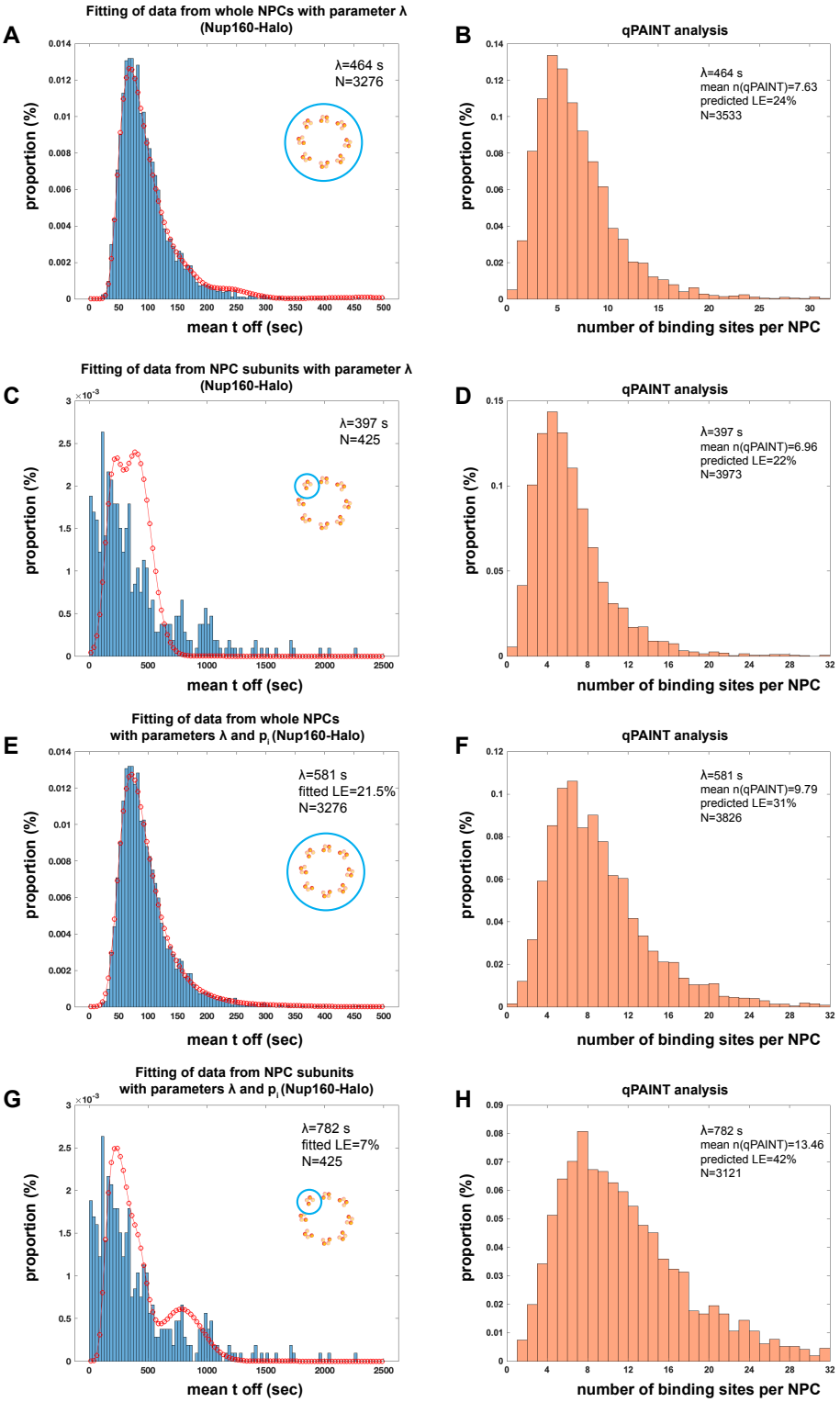


Figure 4.13 **Determining the influx rate for a single binding site by single fitting of Nup160-Halo cumulative mean t_{OFF} obtained in fruit fly follicle cells.** Legend on the next page.

Figure 4.13 *(previous page)* (A) Fitting (the red line) with a single parameter λ of the distribution of cumulative mean t_{OFF} for data of whole NPCs. The probability (p_i) of the number of labelled binding sites was calculated from the effective labelling efficiency (17%). (B) The distribution of the number of binding sites per NPC as calculated with qPAINT using the influx rate determined in (A). (C) Fitting (the red line) with a single parameter λ of the distribution of cumulative mean t_{OFF} for data of NPC subunits. The probability (p_i) of the number of labelled binding sites was calculated from the effective labelling efficiency (17%).(D) The distribution of the number of binding sites per NPC as calculated with qPAINT using the influx rate determined in (C). (E) Fitting (the red line) with parameters λ and (p_i) of the distribution of cumulative mean t_{OFF} for data of whole NPCs. The fitted labelling efficiency (LE) was calculated as 21.5%. (F) The distribution of the number of binding sites per NPC as calculated with qPAINT using the influx rate determined in (E). (G) Fitting (the red line) with parameters λ and (p_i) of the distribution of cumulative mean t_{OFF} for data of NPC subunits. The fitted labelling efficiency (LE) was calculated as 7%. (H) The distribution of the number of binding sites per NPC as calculated with qPAINT using the influx rate determined in (G).

The same analysis was performed on cumulative mean t_{OFF} data from NPC subunits (n=425, manually chosen). λ was measured as 397 s, which translated into influx rate of 0.002518 s^{-1} (Figure 4.13C). The average number of binding sites per NPC was 6.96, which would translate into 22% predicted labelling efficiency (Figure 4.13D).

Again I also calculated λ using the double fit of the data. For the data from the entire NPCs the double fit yielded parameter λ of 581 s (influx rate=0.001221 s^{-1}) and 21.5% fitted labelling efficiency (Figure 4.13E). The average number of binding sites per NPC calculated with this influx rate was 9.79, which would translate into 31% predicted labelling efficiency (Figure 4.13F). For the data from the NPC subunits the double fit yielded similar parameter λ of 782 s (influx rate=0.001278 s^{-1}) and 7% fitted labelling efficiency (Figure 4.13G). The average number of binding sites per NPC calculated with this influx rate was 13.46, which would translate into 42% predicted labelling efficiency (Figure 4.13H).

Altogether, based on the available evidence the best estimate of the influx rate in the fruit fly follicular epithelium is λ=464 s. I used this λ for the quantifications of polarity proteins in Chapter 5. Despite slight discrepancy in predicted labelling efficiency with the ELE, one has to consider that the latter was assessed on the small number of pores that were manually selected.

			Fitting with λ	**qPAINT analysis**		**Fitting with λ and p_i**		**qPAINT analysis**	
Exp. system	**Tagged protein**	**Data**	Exp. LE for fitting	λ	qPAINT result based on λ	predicted LE from qPAINT	λ	Fitted LE	qPAINT result based on λ	predicted LE from qPAINT
U2OS	Nup96-Halo	whole NPC	32%	414s (n=1278)	10.34 (n=6649)	32%	682s (n=1278)	43%	16.65 (n=7606)	52%
		NPC subunit		222s (n=136)	5.77 (n=6386)	18%	863s (n=136)	16%	20.39 (n=5097)	64%
	Nup96-SNAP	whole NPC	21%	692s (n=8567)	9.37 (n=5332)	29%	879s (n=8567)	27%	11.90 (n=5332)	37%
		NPC subunit		972s (n=129)	13.16 (n=5332)	41%	884s (n=129)	18%	12.75 (n=5332)	39%
Drosophila	Nup160-Halo	whole NPC	17%	464s (n=3276)	7.63 (n=3533)	24%	581s (n=3276)	31%	9.79 (n=3826)	31%
		NPC subunit		397s (n=425)	6.96 (n=3973)	22%	782s (n=425)	7%	13.46 (n=3121)	42%

Table 4.1: **Summary of all experimental results presented in this chapter.** qPAINT result denotes the number of binding sites calculated according to qPAINT approach, NPC=nuclear pore complex, LE=labelling efficiency.

4.2.4 Determining the absolute labelling efficiency for polarity proteins

Finally, to quantify the absolute numbers of polarity proteins, their ALE had to be determined. For aPKC-Halo and Par6-Halo the gel band shift could be detected (Figure 4.14A and Figure 4.14B) and the labelled fraction quantified as on average 36.4% (std=7.9%) and 49.6% (std=6.9%), respectively (Figure 4.14D). However, gel band shift in Crumbs-SNAP samples could not be detected and they appeared smeary (Figure 4.14C), presumably due to size of the tagged protein (253 kDa). Numerous attempts to run the gel for different times and voltages did not yield success (data not shown).

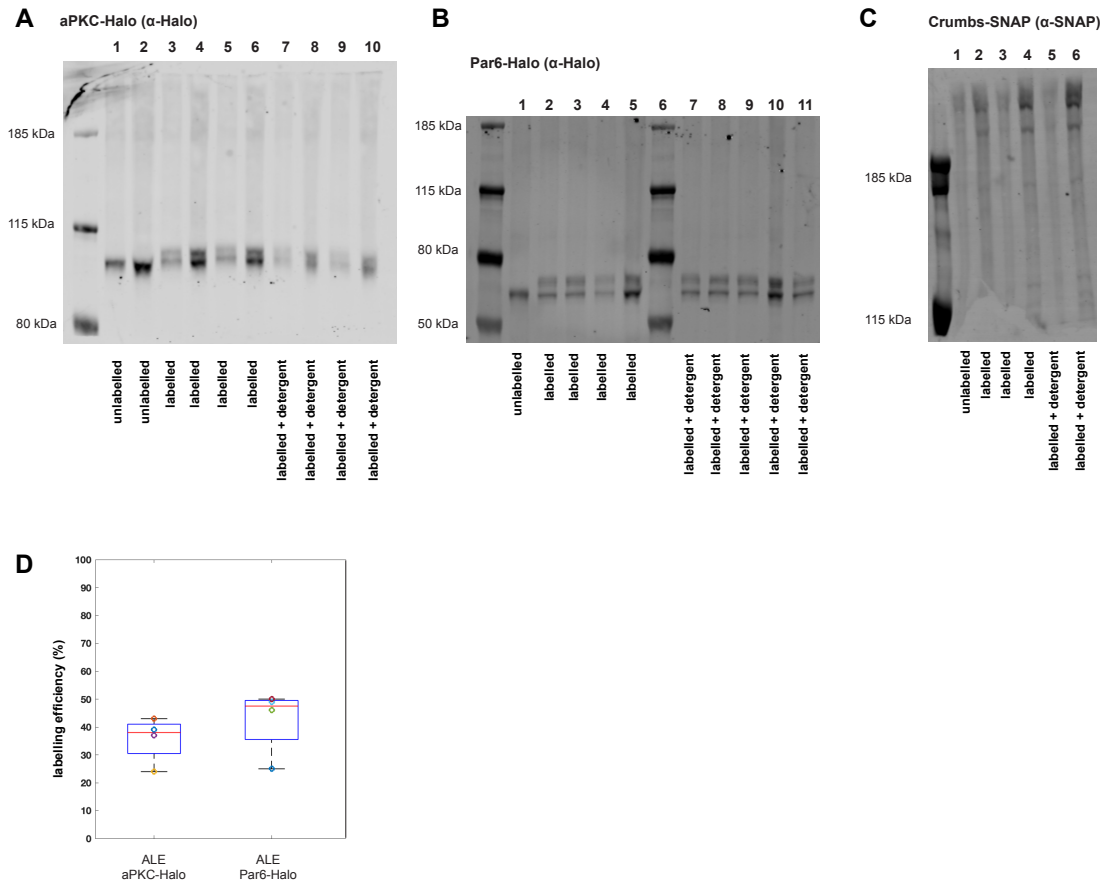


Figure 4.14 **Determining the absolute labelling efficiency for polarity proteins.** Legend on the next page.

Figure 4.14 *(previous page)* (A) A Western blot of a lysate from fixed fruit fly follicle cells expressing aPKC-Halo probed with an antibody against the Halo epitope. Lane 1: a protein lysate from unlabelled follicle cells. Lane 2: a protein lysate (higher concentration) from unlabelled follicle cells. Lane 3-6: a protein lysate (increasing concentration) from labelled follicle cells. Lane 7-10: a protein lysate (increasing concentration) from follicle cells labelled in the presence of 0.5% Tween 20 detergent. (B) A Western blot of a lysate from fixed fruit fly follicle cells expressing Par6-Halo using an antibody against the Halo epitope. Lane 1: a protein lysate from unlabelled follicle cells. Lane 2-5: a protein lysate (increasing concentration) from labelled follicle cells. Lane 6: ladder. Lane 7-11: a protein lysate (increasing concentration) from follicle cells labelled in the presence of 0.5% Tween 20 detergent. (C) A Western blot of a lysate from fixed fruit fly follicle cells expressing Crumbs-SNAP using an antibody against the SNAP epitope. Lane 1: a protein lysate from unlabelled follicle cells. Lane 2-4: a protein lysate (increasing concentration) from labelled follicle cells. Lane 5-6: a protein lysate (increasing concentration) from labelled follicle cells, during labelling the 0.5% Tween 20 detergent was used. Note that the band shift is present but the two bands cannot be distinguished. (D) A plot showing the absolute labelling efficiencies for the aPKC-Halo and Par6-Halo samples. Note that there are no data for the absolute labelling efficiency for Crumbs-SNAP since the band shift could not be distinguished and quantified (see (C)).

4.2.5 Differences in influx rates between the experimental setups

Additionally, I would like to summarise the differences of influx rates between different experimental setups (i.e. DNA origami, cultured cells, tissue sample) and to highlight the necessity of experimental work presented in this chapter. In DNA origami the influx rate was on average 1 binding event every 358 s ($c_{\text{imager oligo}}$=2 nM), while in U2OS cells expressing Nup96-Halo the influx rate was on average 1 binding event every 414 s ($c_{\text{imager oligo}}$=2 nM). In follicle cells I used higher concentration of the imager oligo ($c_{\text{imager oligo}}$=4 nM) the influx rate was on average 1 binding event every 464 s. Only a small decrease in the observed influx rate between cells and tissue, despite two times higher concentration of the imager oligo, suggests that diffusion in tissue is slower. Especially since higher concentration of the imager oligo should be reflected in more frequent binding events. Further, the influx rate in the U2OS cells expressing Nup96-SNAP that were imaged using 1 nM concentration of the imager oligo was 1 binding event every 692 s. Altogether, these results suggest the importance of influx rate calibration for the environment where molecular target of interest reside.

4.3 Discussion

Here, an approach to determine the influx rate for a single binding site in tissue is presented. I have shown that it differs from that obtained in vitro (e.g. DNA origami, cultured cells). Hence the influx rate must be determined in the environment where the protein of interest resides. The nuclear pore complex (NPC) is an attractive biological structure from which to determine the influx rate for a single binding site because of the known nucleoporin protein stoichiometry. Using simple mathematical fitting the influx rate can be estimated.

Counting molecules was part of the original qPAINT work presented by Jungmann and co-workers (Jungmann et al. 2016). The authors used a single isolated target protein to calibrate the influx rate (Jungmann et al. 2016). In the case of NPCs, they labelled Nup98 using a monoclonal primary antibody directly conjugated to a docking oligo. Then they used single isolated Nup98 protein clusters (or NPC subunits) as a calibration for the imager oligo influx rate. They state that the antibody labelling is potentially imperfect and therefore each NPC shows different number of individual subunits. However, they do not address the existence of multiple protein copies per NPC subunit. Instead they quantified the number of subunits per NPC. This is most likely the reason they achieved 95% accuracy and 84% precision. The single isolated NPC subunit that served as a calibration standard probably had two binding sites labelled, which in the context of counting the NPC subunits should not yield significant errors. However, this approach would yield an error in the case of counting the number of labelled binding sites per NPC, since the influx rate was not calibrated for a single binding site.

For an in vivo example they applied the same approach to quantify the number of single Bruchpilot proteins in the fruit fly neuromuscular junction. They first used monoclonal primary antibodies against Bruchpilot protein and then a docking oligo-conjugated secondary antibodies for labelling. Similarly, they used single protein targets to calibrate the influx rate. Again, two sources of errors with the quantification stem from this approach. First, multiple secondary antibodies bound per primary antibody. And second, the single protein targets used for the calibration of the influx rate are presumably single, but could be also double, for example.

In another study, Soeller and co-workers expanded this approach when quantifying ryanodine receptors in rat myocyte (Jayasinghe et al. 2018). They quantified the mean t_{OFF} of numerous small clusters, presumably single ryanodine receptor proteins. The distribution

of mean t_{OFF} exhibited characteristic equidistant peaks suggesting the analysed clusters contained one, two, three, etc. ryanodine receptor proteins. However, they reasoned that the largest peak (exhibiting the longest mean t_{OFF}) belonged to a single target protein and used its influx rate to quantify the super-resolution acquisitions.

Both examples above were ground-breaking in the sense that they were able to obtain the influx rate for single binding sites. However, their success is, in part, a result of working in more forgiving systems. In the first example, they were only counting the number of NPC subunits. In the second example, the influx rate could be determined by sampling the smallest protein clusters and determining the longest mean dark time. However, it is more challenging to address this problem if the protein of interest is not present as a monomer or if the latter is difficult to distinguish from the non-specific (background) signal, as is often the case when using thick tissue samples, like the fruit fly egg chamber.

Here I extracted this information using the mathematical approach, fitting the distribution of mean t_{OFF} times with a single parameter λ. An important conclusion from comparing one vs. two parameter fitting is that a two parameter fit usually yields more erratic estimate of the value λ. This is counterintuitive as fitting with more parameters usually means a better fit, but in this case this does not mean a more accurate one. This was evident from the DNA origami simulated data where a two-parameter fit resulted in suggested labelling efficiency that was not calculated correctly. Experimentally, this was especially apparent with the double fit of the data from Nup96-Halo NPC subunits where the calculated number of the binding sites resulted in a 64% predicted labelling efficiency. Based on the probabilistic fit, this should mean that 9 out of 10 NPCs should have all 8 subunits labelled, which was clearly not the case. While one could explain this with a low sample size (n=136), double fitting of the data from the whole NPC (n=1278) resulted in 52% predicted labelling efficiency. This should still result in majority of pores having labelled 7 or 8 subunits, which was again not the case.

The conclusion I would like to make is that binding kinetics data are probably of noisy nature per se. This could be explained by the multistep process that produces the dataset for fitting. Each step probably results in a small error. For example blinks that were not fitted and localised result in a longer mean t_{OFF}, while automated image segmentation of the NPCs selects some non-specific binding events, etc. While fits of the simulated DNA-PAINT images, where non-specific background is not present, were better they were still not perfect. It would be interesting to see how the fitting would perform in a

completely noise-less simulation without the outliers. Nevertheless, the single fit approach presented here is robust enough to determine λ value and the big sample size improves the accuracy of the estimation.

One way to circumvent the possible errors with estimating the value of λ value with the fitting approach would be to perform the binding kinetics on Gle1 protein. It is one of the proteins in the NPC subunits of the cytoplasmic ring (Ori et al. 2013). Importantly, it is present in only one copy per subunit. Having one binding site per subunit means that one would just need to analyse the binding kinetics and determine the mean t_{OFF} of the subunit area. Gle1 is also relatively small protein (77 kDa), which should simplify assaying the ALE with the gel band shift assay.

An important question that arises is if the influx rate determined using NPC can be also used to quantify the transmembrane proteins, where the tag is present on the extracellular side. The diffusion rate of molecules outside the cell might be different which would change the influx rate as well.

A second novelty presented here that adds to the general method of quantitative DNA-PAINT is determining the absolute labelling efficiency using the gel band shift assay. Since the molecular target population is usually not fully labelled, this information is crucial to quantify the absolute number of protein molecules without undercounting. Consequently further biological information can be deduced, e.g. the cytoplasmic protein concentration versus the cortical protein concentration. Moreover, the results suggest that the labelling efficiency is protein-dependent and not tag-dependent. However, it was reported before that there is some inherent difference between Halo- and SNAP-tagged proteins (Erdmann et al. 2019). The incomplete labelling could be caused by various reasons. Firstly, the protein microenvironment could affect ligand accessibility. For example, one could expect that a transmembrane protein with a tag positioned in the extracellular domain will be more accessible for the ligand than a nuclear protein. Secondly, the sample preparation might be also important to achieve optimal labelling and the protocol optimisation was not fully explored in this work. Nevertheless, the range of labelling efficiency for Halo- and SNAP-tagged proteins quantified here follows a recent report in the context of STORM imaging (Thevathasan et al. 2019).

The gel band shift assay approach to quantify the absolute labelling efficiency has a number of experimental subtleties. Due to the small difference (circa 3kDa) in molecular weight of

the labelled protein fraction, it is more difficult to determine the labelling efficiency of high-molecular weight proteins (>130kDa, excluding the tag) since the gel has to be run for longer time creating indistinct band shifts. This could be overcome by adding a TEV protease cleavage site in the protein sequence upstream of the Halo or SNAP tag. The protein sample would then be treated with the protease resulting a much smaller peptide. The peptide with the Halo-tag would be then used to assay the ALE.

I would also like to note that a loading control should be used in future experiments to indicate equal loading of samples across the wells and thus to make quantifications more accurate.

4.4 Perspectives

In this chapter I presented the pipeline for obtaining two crucial parameters when counting the number of molecules in tissue sample using DNA-PAINT approach in vivo: the influx rate for a single binding site and the labelling efficiency of the investigated polarity proteins.

In the following chapter I describe how I used the influx rate to calculate the absolute number of polarity proteins along the cell junctions and in the cytoplasm. I elaborate on the initial observation that polarity proteins form clusters and try to describe this clustering in a quantitative manner.

4.5 Acknowledgment of contributions

The Nup160-Halo and Nup160-SNAP transgenic fly lines were created by Jenny Richens and Amandine Palandri. The automated analysis of effective labelling efficiency pipeline was developed in collaboration with Richard Butler from the Gurdon Imaging Facility. The mathematical fitting pipeline was developed in collaboration with David Jordan. Vivien Tsang provided assistance in setting up the gel band shift assay protocol and performing the Western blots results used in this thesis. The Nup96-Halo and Nup96-SNAP cell lines were provided by Jonas Ries from European Molecular Biology Laboratory in Heidelberg.

Chapter 5

Characteristics of polarity proteins spatial organisation *in vivo*

5.1 Introduction

Many cytosolic proteins can organise in space and time to form membrane-free or associated high-ordered assemblies. Examples include proteins organising into subcellular structures with a known architecture, like tubulin that organises into the microtubules (Desai and Mitchison 1997) or nucleoporins that build the nuclear pore complexes (Lutzmann et al. 2002). On the other hand, many of the proteins compartmentalize into much less defined molecular assemblies, like the MEG proteins building RNA granules (Wang et al. 2014) or the clustering of RNA polymerase II (Cisse et al. 2013).

Nevertheless, for many proteins, like polarity proteins, it is still not known if they are able to form oligomeric molecular assemblies besides the isolated examples of worm zygote during polarity establishment (Lang and Munro 2017) and reconstitution of Par3 clustering in cultured fruit fly S2 cells (Kono et al. 2019). However, this does not mean that these proteins do not organise into distinct spatial compartments with specific functions also in other systems, like fruit fly epithelial cells. One of the most common patterning is molecular clustering (Garcia-Parajo et al. 2014). In this chapter, I describe the experiments carried out to determine if the characteristics of the spatial organisation of polarity proteins can be mathematically modelled. For this I collaborated with computer scientist Leila Muresan (Cambridge Advanced Imaging Centre).

5.1.1 Clustering as a major mode of spatial molecular patterning

Clusters are defined as assemblies of molecules, which bind to each other in a dense structure and can be ordered or not as previously defined (Recouvreux and Lenne 2016). Clustering is often defined as a reversible process and thus differs from aggregation that is an irreversible process (Weber et al. 2019).

Clustering has been well described for the lipid-anchored membrane proteins. Some of the most well understood ones are glycosylphosphatidyl (GPI)-anchored proteins that can be involved in a wide variety of physiological roles from cell signalling to cell adhesion (Sharma et al. 2004). Another well documented protein group exhibiting clustering is the Ras superfamily of small GTPases. Rac1 is mobile when in monomers but is immobilised upon clustering. The Rac1 clusters have diameters of approximately 200 nm and contain up to 50 proteins (Remorino et al. 2017). Other well studied membrane proteins are linkers for activation of T cells (LAT) that form signalling assemblies for regulation of the actin cytoskeleton in T cells (Su et al. 2016). Extensive research has been done on receptor signalling clusters in the immune and neuronal synapses (Dustin and Groves 2012; Hartman and Groves 2011).

Only recently the field has been moving also towards protein assemblies that are not associated with the plasma membrane. Some of the best characterised examples include RNA-protein granules in the nucleus (Mao et al. 2011), as well as stress granules and germ granules in the cytoplasm (Seydoux 2018). The discovery about their liquid-like properties has opened an avenue of protein phase separation research in cell biology (reviewed in (Banani et al. 2017)). This is a physical process that results when a mixture of protein A and B spontaneously separates into two phases A and B that then stably coexist. Separation can happen due to increasing concentration of one of the components, increasing valency or affinity of components, or decreasing intrinsic solubility of one of the components (reviewed in (Banani et al. 2017)).

In the context of cell polarity protein-protein phase separation has been described only for Numb and Pon proteins during neuroblast asymmetric division in the fruit fly (Shan et al. 2018). This observed Numb and Pon protein condensation is thought to happen via multivalent interactions. As I already described in Chapter 1, a brief clustering process of polarity proteins has been well studied in the worm zygote during polarity

establishment. Importantly, no liquid-like characteristics have been described for these assemblies. Moreover, it seems that protein clustering does not play a role during polarity maintenance (except for CHIN1 protein), at least not in this system.

5.1.2 Importance of counting the molecules

Some physical characteristics of the clusters such as their shape and area or volume can be quite easily obtained from qualitative super-resolution imaging while other parameters, like the number of proteins that a single cluster contains are relatively difficult to quantify. The latter demands not only good quality super-resolution images but a robust counting approach as well. Because we operate with snapshots of a steady-state system we cannot observe how clusters dynamically emerge and behave. However, according to extensive research in the field of statistical physics these snapshots also contain a wealth of information on cluster dynamics (Peruani et al. 2006; Abney et al. 1987; Braun et al. 1987).

The cluster size distribution is the main physical measure necessary to develop models about cluster formation and maintenance. Importantly, these models then help to predict how would clusters change upon varying protein concentration, trafficking and binding rate (Truong Quang and Lenne 2014). The cluster size distribution can be usually described by different statistical distributions and this can be then used to make inferences about different modes of cluster formation.

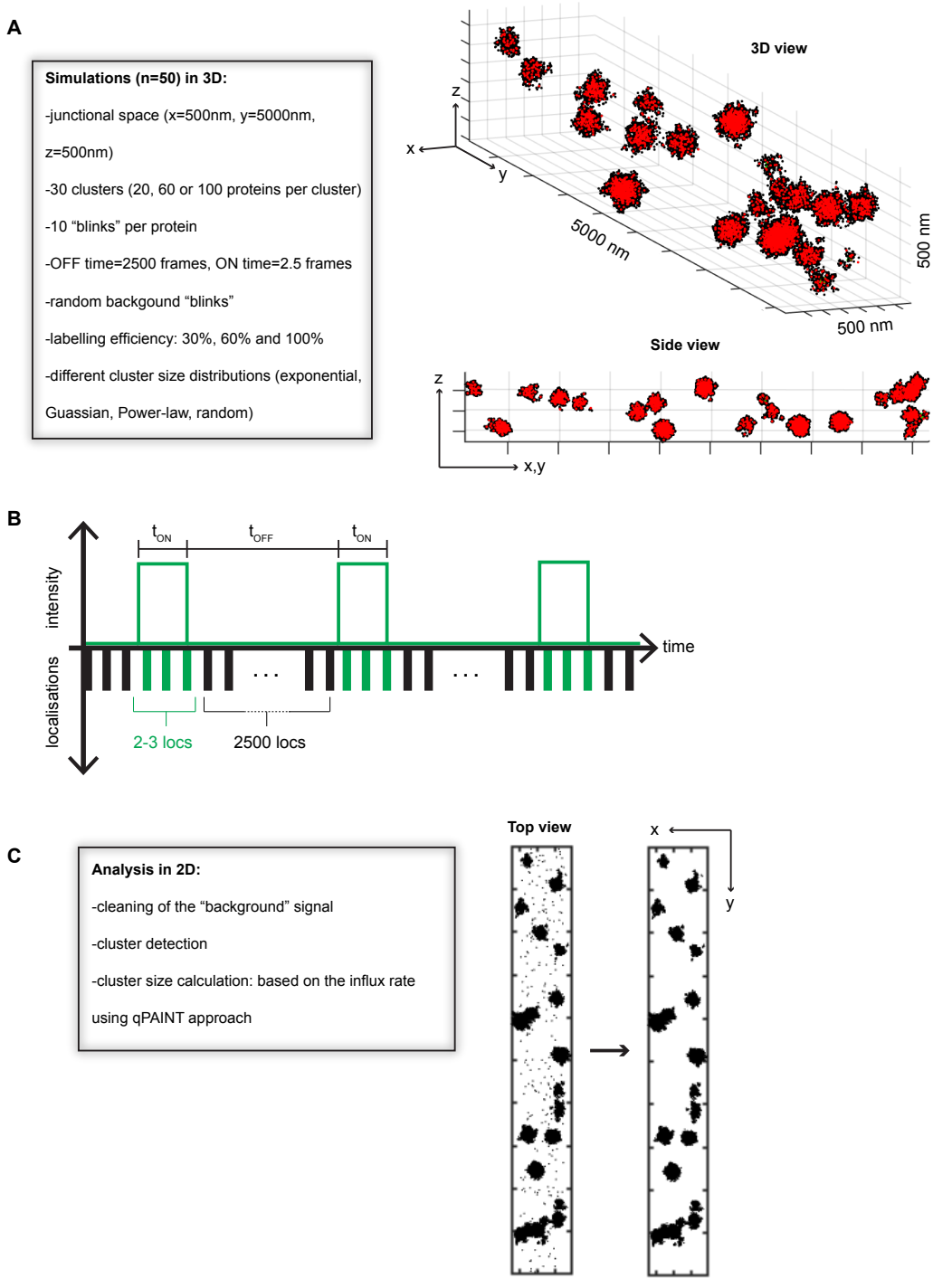


Figure 5.1 **Work-flow for computer simulations of protein spatial organisations and their analysis.** Legend on next page.

Figure 5.1 *(previous page)* (A) Left: a description of the main parameters for computer simulations of different cluster size distributions in 3D space. Right: an example of a simulation result with a cluster size distribution approximated with an exponential function. Red dots represent proteins and black dots represent localisations due to "blinking" (binding of the imager oligo to the docking oligo). (B) A temporal trace of imager oligo hybridisation events to the docking oligo and how this relates to the number of localisations in the computer simulations. (C) Left: Description of the main steps in the 2D analysis of simulated data. Right: an example of the simulated result shown in (A) viewed from the top as a maximum projection in the z-axis, before and after cleaning the "background" signal.

For example, it was found that bacterial chemotaxis receptors follow an exponential distribution. This can be explained by stochastic self-assembly – receptors would diffuse within the membrane and collide with other receptors (Greenfield et al. 2009). In the case of E-cadherin in epithelial cells of the fruit fly embryo, the clusters follow a power law distribution with an exponential cut-off. Modelling suggested that E-cadherin clusters undergo dynamics fusion and fission events, however upon reaching a certain size they are endocytosed (Truong Quang et al. 2013). Importantly, the density of proteins can change their cluster size distribution, as it was observed for nephrin clusters. At lower densities the distribution is exponential, while at higher densities it follows a power law (Banjade and Rosen 2014).

This chapter proposes an approach to quantitatively describe the cluster size distribution of apical polarity proteins in follicle cells of the fruit fly. Before we analysed biological data (i.e. super-resolution images of proteins), we firstly validated our analysis approaches using simulated data. I describe this experimental design in the next section.

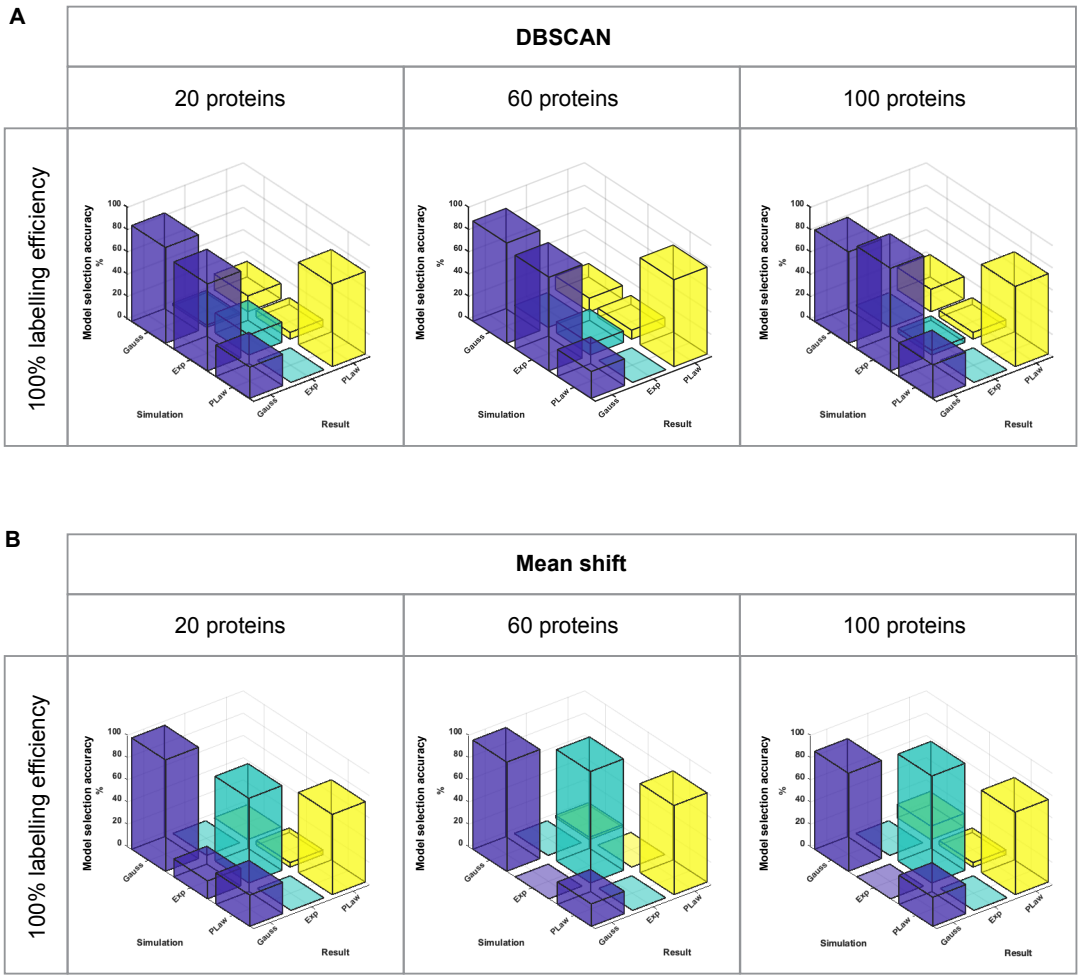


Figure 5.2 **Comparing DBSCAN and mean-shift algorithm for cluster identification.** Legend on next page.

Figure 5.2 *(previous page)* (A) Resulting model selection from the DBSCAN-based cluster analysis. The minimum size between clusters was set to 0 (crowding allowed). In the simulations the labelling efficiency was 100%. (B) Resulting model selection from the mean-shift-based cluster analysis. The minimum size between clusters was set to 0 (crowding allowed). In the simulations the labelling efficiency was 100%.

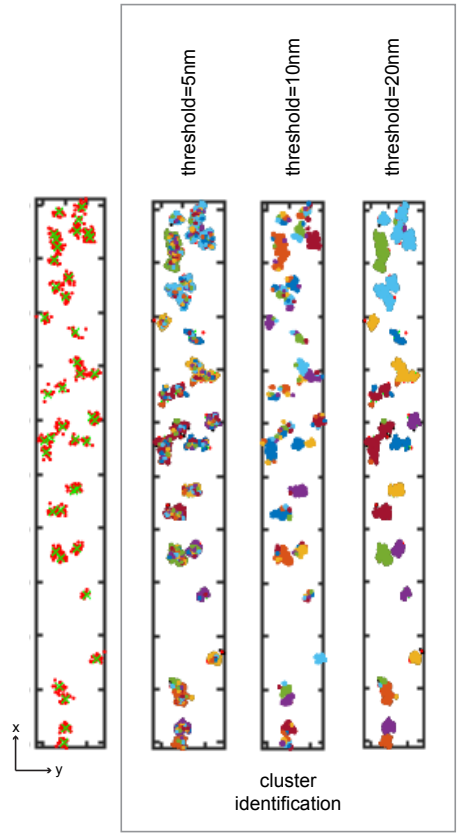


Figure 5.3 **Effect of threshold in DBSCAN on cluster identification.** Left: a top view of an example of simulated data on which DBSCAN cluster identification was performed (red points: protein molecules, green points: cluster centres). Right: the simulated data (shown on left) analysed using DBSCAN with different threshold values. The identified clusters are colour-coded.

5.2 Experimental design

5.2.1 The theory behind computer simulations

Computer simulations have been extensively used as a tool in modelling the spatial organisation of proteins, especially for those in the membranes (Lindahl and Sansom 2008;

Khalili-Araghi et al. 2009; Sieber et al. 2007; Grossfield et al. 2007). They are an excellent approach for validation of algorithms for cluster detection since the ground truth is known. The main motivation behind these verifications of the simulated data was to be more confident when interpreting the biological data. To this end, we simulated cell junctions with different molecular spatial organisations in terms of cluster size distributions to verify the accuracy of our cluster detection approaches (Figure 5.1A).

The computer simulations were approached as a three-dimensional spatio-temporal point process in a confined space (a synthetic cell junction) $W = [500] \times [4000] \times [700]$ nm. The protein cluster centres were defined as:

$$c_i = (x_i, y_i, z_i), i \in \{1, \ldots, N\} \tag{5.1}$$

The number of protein clusters was:

$$n \sim Poi(N), N = 30 \tag{5.2}$$

The spatial distribution of the cluster centres was one of the following two models: completely spatially random process and hard core process (where no two clusters are closer than distance $T = 100$ nm. Proteins in cluster i were denoted: $\{p_{ij} = (x_{ij}, y_{ij}, z_{ij}), j \in \{1, \ldots, n_i\}\}$. The location of proteins was assumed Gaussian distributed (with $\sigma = 30$) around the cluster centers c_i.The model described above is known as a Neyman-Scott process, more specifically a (modified) Thomas process and here it aims to describe the biological aspect of the protein clustering.

The number of proteins in a cluster (cardinality) was distributed according to one of the following models (with mean $M \in \{20, 60, 100\}$):

- exponential distribution

$$m_i \sim \mathcal{E}(M), \quad p(m_i) = M \exp(-M m_i) \tag{5.3}$$

- Gaussian (normal) distribution

$$m_i \sim \mathcal{N}(M, \sigma), \quad p(m_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(m_i - M)^2}{2\sigma^2}\right) \tag{5.4}$$

- power law distribution

$$m_i \sim \mathcal{PL}(M), \quad p(m_i) = \alpha \cdot m_i^{-1/\alpha - 1}, \quad \alpha = 1.5 \tag{5.5}$$

However, due to DNA-PAINT approach the kinetics of the fluorescence signal induces an additional level of complexity. Firstly, not every protein was labelled. In order to simulate labelling efficiency, each protein was marked as a fluorescence signal emitter with a pre-defined probability. Secondly, a labelled protein will exhibit ON/OFF blinking pattern (simulating the binding and unbinding of the fluorescently labelled imager oligo). The blinks pertaining to protein $ij : b_{ijk} = (x_{ijk}, y_{ijk}, z_{ijk}), j \in 1, \ldots, m_i$. Spatially, the location of blinks represents a second level of spatial Gaussian clustering, $\sigma_{loc} = 10$ nm, around the position of each labelled protein p_{ij}. σ_{loc} reflects the localisation precision of the super-resolution imaging. This complex process is know as a double clustering (or hierarchically clustered) process.

For the binding kinetics the following mean dark and bright times were used. Mean dark time = 2500 frames, mean bright time = 2.5 frames (Figure 5.1B). The length of the simualted experiments was 10000 frames. Therefore, each protein was then assigned approximately 10 localisations per protein molecule, which corresponded on average to 4 binding events of the imager strand per binding site (i.e. per protein molecule).

Note that the position of the fluorescent proteins p_{ij} or that of cluster centres c_i is not known (not observed). Only the localisations b_{ijk} can be estimated from the image measurements. Moreover, given the imaging limitations, only a 2D projection of the localisations is available (along the z axis).

5.2.2 The theory behind the model selection

We used Bayesian information criterion (BIC) to select the hypothesis that would be the best explanation for the cluster size distribution. Bayesian model selection makes use of the Bayesian theorem to choose the most plausible hypothesis. In our case the hypotheses M_k are the different cluster size distributions (Gaussian, exponential, power-law, random) with different parameter spaces Θ_{Mk} (Bernardo and Smith 2008). The so called posterior probability (after the evidence of the data has been taken into account) is derived from a prior probability (a prior knowledge on the considered models) and a likelihood function (the probability that the data was generated by the model). By Bayes theorem the following holds:

$$P(M_k \mid D) = \frac{P(D \mid M_k)\, P(M_k)}{P(D)} \tag{5.6}$$

where $P(M|E)$ is the posterior probability of the model (M_k) given the data (D), $P(D|M_k)$ is probability to observe the data given the model, $P(M_k)$ is the probability of the model prior and $P(D)$ is the probability of the data. The probability to observe the data given the model is also called likelihood. In our approach we calculated the negative-log likelihood for each of the models and then selected the model that had the lowest negative-log likelihood ratio (see Appendix J for more details).

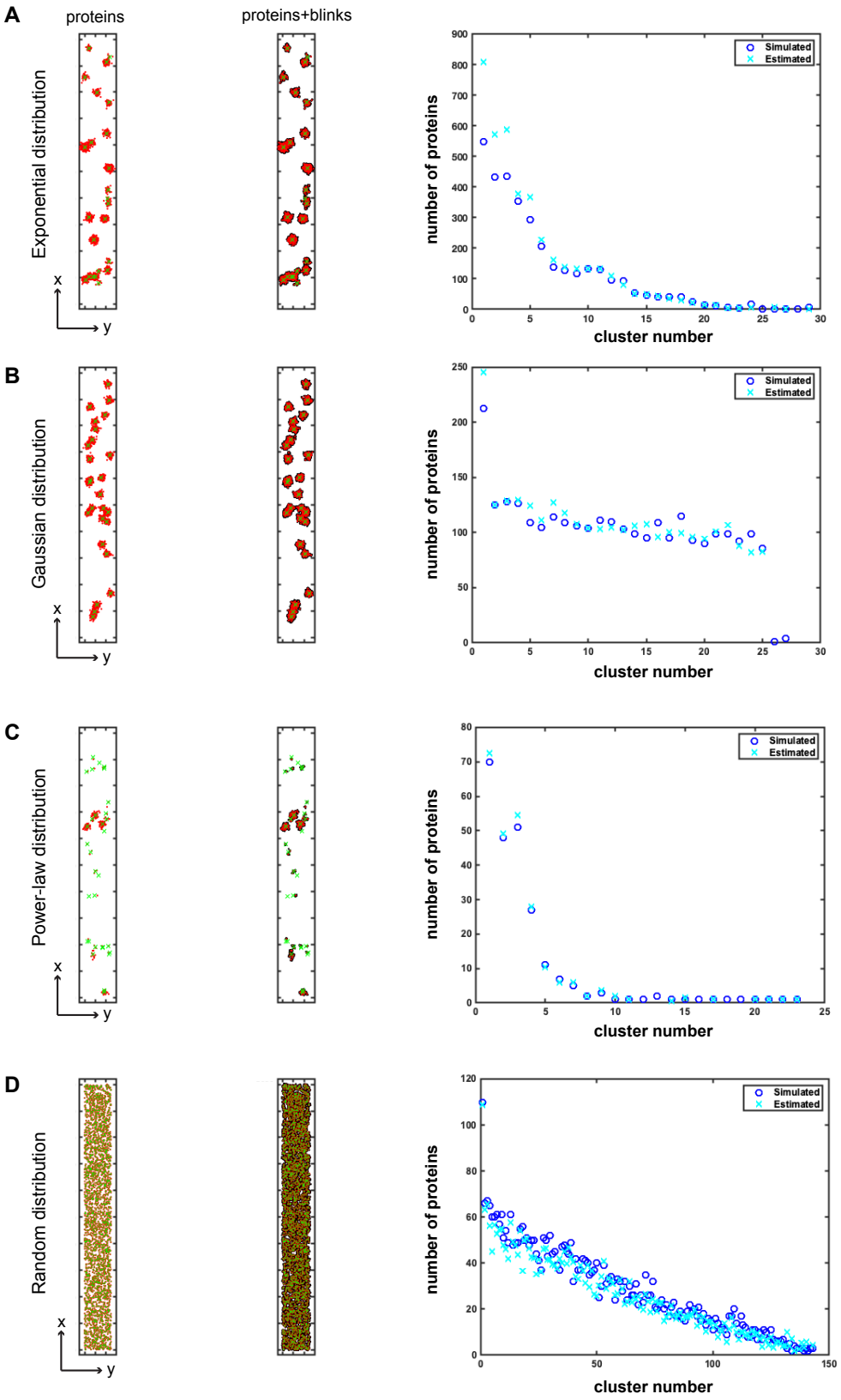


Figure 5.4 **Examples of different protein spatial organisations from computer simulations.** Legend on next page.

Figure 5.4 *(previous page)* Examples of (A) exponential, (B) Gaussian, (C) power-law, and (D) random distribution of protein cluster size (number of molecules). In each panel a top view of the junction is shown with the protein clusters (left), with the protein clusters and underlying localisations due to the blinking simulation (middle), the correlation between the simulated and the estimated cluster size (right). Green crosses mark the cluster centres. In the simulations shown the labelling efficiency was 100%.

5.2.3 Validating DBSCAN and mean-shift algorithm for cluster identification

For cluster identification, the two dimensional top projection of the three-dimensional simulated junction was considered. Just as for real data analysis two dimensional super-resolution images of three dimensional slices are used. Then the junction was cleaned of the "background" signal by removing all single localisations and all clusters containing 10 or fewer localisations (Figure 5.1C).

Next we wanted to identify the best approach for accurate cluster detection. Density-based spatial clustering of applications with noise (DBSCAN) and the mean-shift algorithm are two widely used clustering algorithms. DBSCAN is traditionally used as the best approach when doing cluster analysis (Ester et al. 1996). DBSCAN relies on two parameters: ϵ that specifies how close points should be to each other to be considered a part of a cluster and m that the minimum number of points to form a dense region. Its advantage is that it can detect clusters of arbitrary shapes, however the strong dependence on the two parameters makes it difficult to separate close by (overlapping) clusters. The mean shift algorithm works by finding a local maxima. It uses kernel density estimation where the bandwidth of the kernel is the only free parameter and defined by the user (Fukunaga and Hostetler 1975; Comaniciu and Meer 2002) .

To rigorously test both algorithms, the clusters positions were simulated using completely spatially random processes, meaning that they were allowed to have arbitrarily small inter-cluster distances as it would occur by pure chance. Additionally, we used three different models (Gaussian, exponential, and power law) of cluster size distribution and different average numbers of proteins per cluster (20, 60, and 100).

Both algorithms performed reasonably well. However, while DBSCAN did not have problems identifying power-law distributed clusters, it had problems identifying cluster

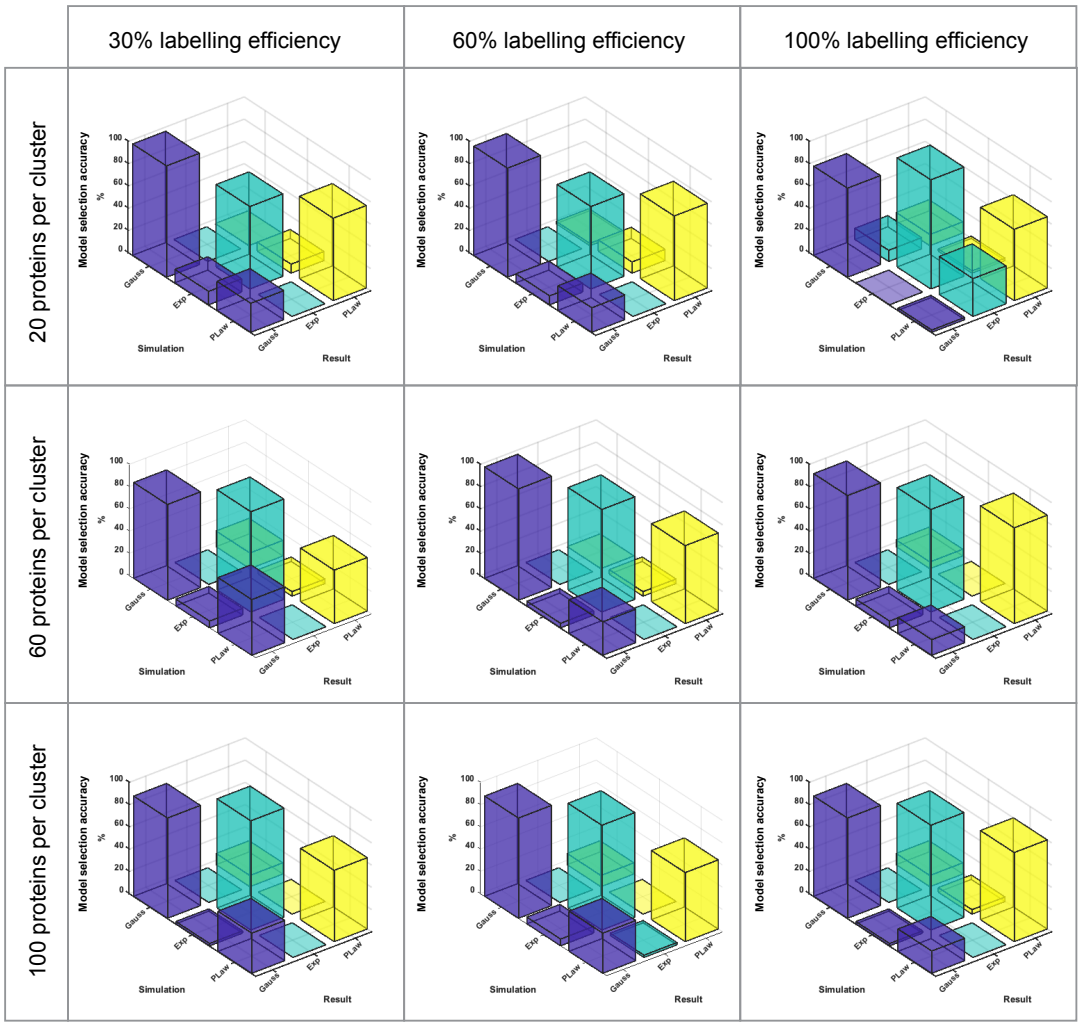


Figure 5.5 **Effect of labelling efficiency and number of proteins per cluster on the model selection of the cluster size distribution.** Cluster analysis using the mean-shift algorithm was performed on simulated data (n=50 for each parameter). Different columns reflect different labelling efficiencies and different rows reflect different average numbers of proteins per cluster.

sizes that were exponentially distributed (Figure 5.2A). This was not the case for the mean shift algorithm, which successfully identified clusters from all three models of molecular distributions (Figure 5.2B).

A visual inspection of DBSCAN classified clusters using different thresholds revealed that while low thresholds find multiple clusters within a single simulated cluster, big thresholds merge multiple simulated clusters into a single one (Figure 5.3). This suggests that DBSCAN seems more sensitive with respect to the threshold parameter, which is difficult to tune. On the other hand, the mean shift algorithm with 50 nm bandwidth did not

exhibit this behaviour and seemed more robust in identifying clusters generated by the double-cluster process described above (Figure 5.4).

5.2.4 Robustness of the mean shift algorithm

Having established the mean-shift algorithm with 50 nm for the size of bandwidth as a best approach to detect clusters, we wanted next to test how robust is this algorithm upon changing different simulations parameters.

I tested how the accuracy of model selection changes upon different number of proteins per cluster (20, 60 or 100) and varying the labelling efficiency (30%, 60% or 100%). In this case clusters were simulated using a hard core process (no two clusters are closer than 100 nm).

Increasing the number of proteins per cluster or decreasing the labelling efficiency did not significantly change the accuracy of model selection. Power-law distributions were misclassified as Gaussian distributions (in less than 10% of cases), but it seems that this discrepancy was lower upon increasing the labelling efficiency (Figure 5.5).

5.3 Results

5.3.1 Validation of protein clustering in biological data

The first main challenge when investigating the spatial properties of molecular organisation is to validate if the molecular clustering is real. This is because a single binding site is visualised multiple times (due to multiple binding events of the imager strand) and each fitted localisation has its own localisation precision error. Therefore, each binding site will be represented with a cluster of localisations (with the number of localisations positively correlating with the acquisition time length). This was demonstrated in DNA origami experiments, where single binding sites were imaged. The majority of localisations from a single binding site produced a cluster of localisations. These localisations were positioned between 5 and 20 nm from the cluster centre (Figure 5.6).

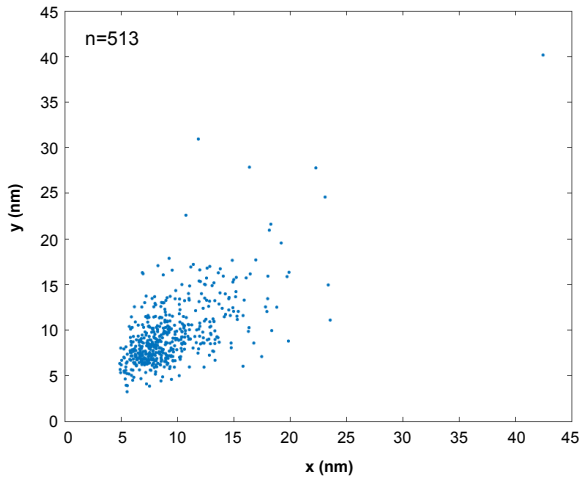


Figure 5.6 **A single binding site produces a cluster of localisations around its centre.** Each dot represents an average spatial spread of all localisations events from the centre of a detected cluster for 513 analysed single binding sites in the DNA origami experiments.

To determine that clustering does not arise only due to multiple binding events (blinking) but also due to underlying protein clustering I used pair correlation function (PCF) statistics. In PCF the probability to find a neighbouring localisation at distance r is quantified for each localisation (Gavagnin et al. 2018). This is then compared with the uniform distribution where the probability is 1. Next the experimental PCF curve is fitted with a model for multiple blinking or for multiple blinking and real clustering (Figure 5.7A) (see Appendix G for more details).

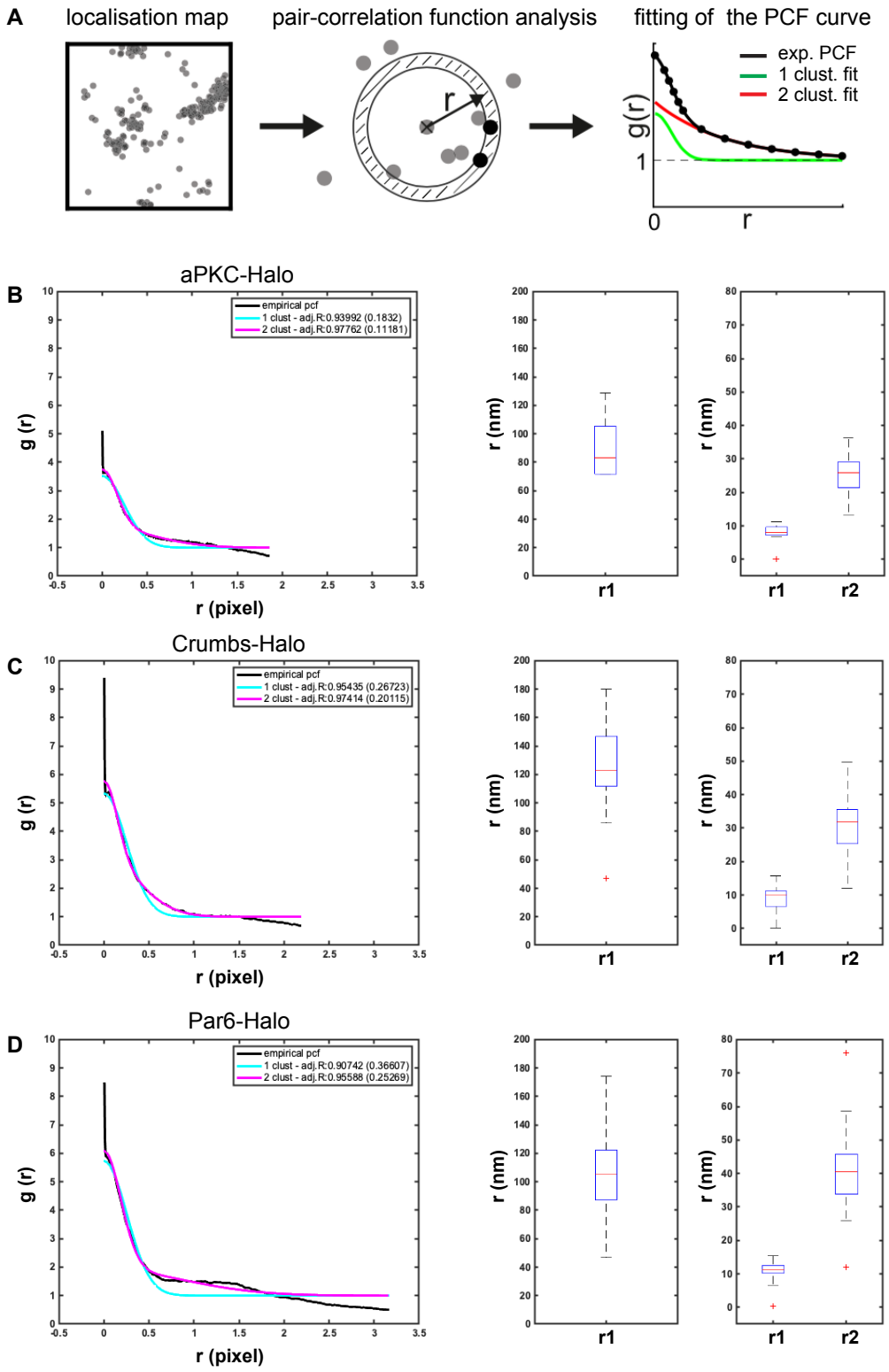


Figure 5.7 **Validation of real clustering of biological data by pair-correlation function (PCF) analysis.** Legend on next page.

Figure 5.7 *(previous page)* (A) A schematic showing the principle behind PCF analysis. The probability of finding a neighbouring localisation at distance r is plotted and compared to a uniform distribution. The experimental PCF is then plotted with the model for a single level of clustering (just due to blinking) and for the double level of clustering (for blinking and protein clustering). Adapted from (Baumgart et al. 2018). (B) Left: An example of a PCF plot for aPKC-Halo along a junction in follicle epithelial cells. Right: Diameter of r1 and r2 from the double fit of the PCF analysis of aPKC-Halo from 5 different junctions. (B) Left: An example of a PCF plot for Crumbs-Halo along a junction in follicle epithelial cells. Right: Diameter of r1 and r2 from the double fit of the PCF analysis of Crumbs-Halo from 5 different junctions. (C) Left: An example of a PCF plot for Par6-Halo along a junction in follicle epithelial cells. Right: Diameter of r1 and r2 from the double fit of the PCF analysis of Par6-Halo from 5 different junctions.

PCF analysis was performed on the super-resolution images of 5 junctions for aPKC-Halo, Crumbs-Halo and Par6-Halo, respectively (Figure 5.7B-D). For all three PCF curves the fit for two layers of clustering was better than a fit where only clustering due to blinking is considered. Interestingly the average radius of the first cluster peak was around 10nm with small data dispersion for all three investigated proteins. This fits with the previously observed data in DNA origami experiments that a single binding sites results in cluster of localisations that are spatially distributed between 5 and 20nm from the cluster centre (Figure 5.6). Interestingly, the average radius of the second cluster peak was bigger with data much more dispersed (Figure 5.7B-D) suggesting that the protein clusters have a range of different sizes (in terms of the area).

These results suggest that there are two levels of clustering of localisations in the super-resolution images of apical polarity proteins. The first due to multiple binding events of the imager oligo and the second due to protein clustering.

5.3.2 Characterising molecular organisation of the polarity proteins

Having established that the protein clustering of polarity proteins is not due to multiple binding events of the imager oligo, I then wanted to briefly investigate the molecular organisation of the polarity proteins on some preliminary super-resolution data acquisitions. To calculate the absolute number of molecules in the clusters and along the cell junctions in this chapter section, I used the influx rate of the imager oligo that was determined from the calculations in the Chapter 4.

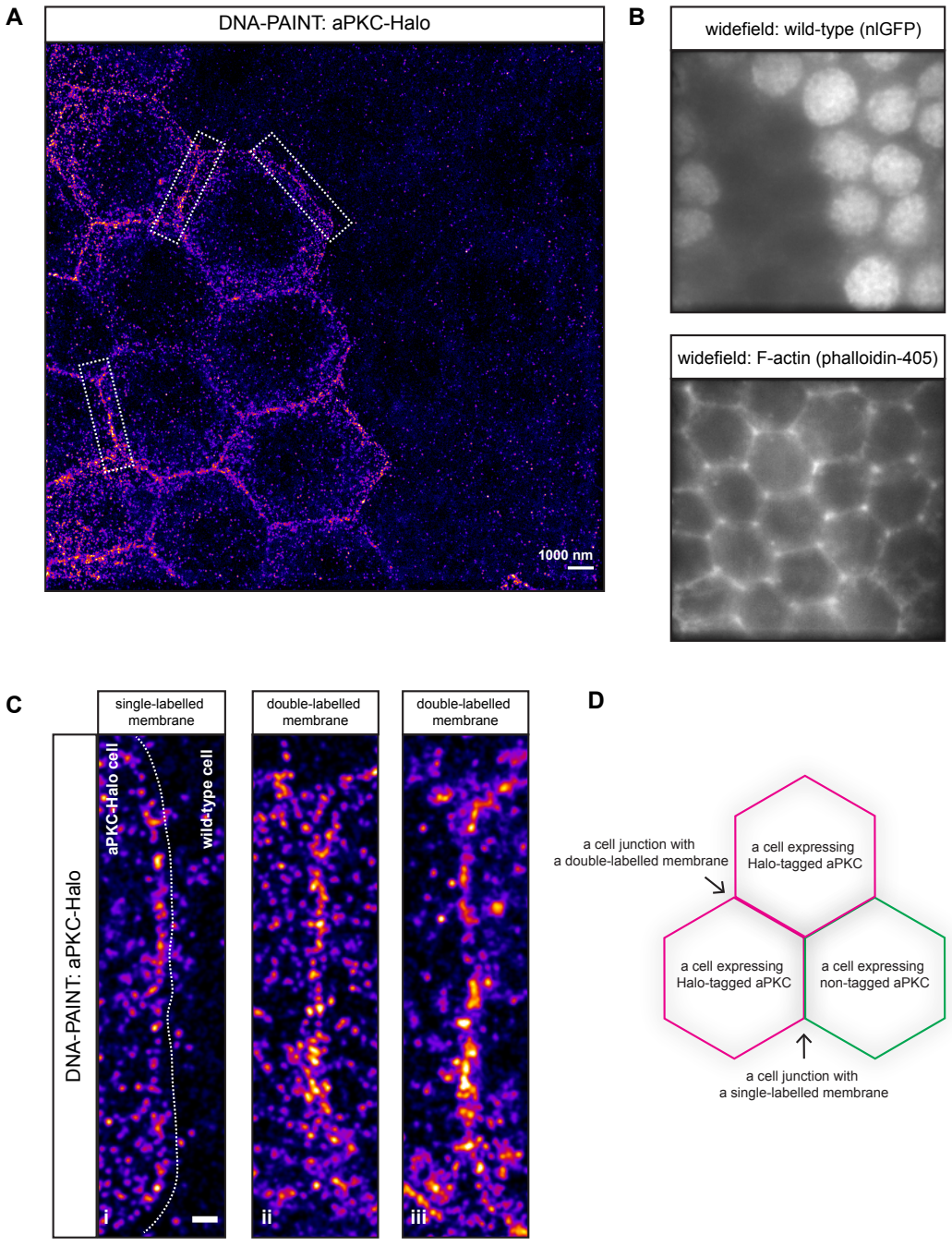


Figure 5.8 **A super-resolution image of aPKC-Halo using DNA-PAINT.** Legend on next page.

Figure 5.8 *(previous page)* (A) An example of super-resolution image showing aPKC-Halo next to wild-type clone cells. (B) Top: the position of the wild-type cells (nuclear signal) relative to the cells expressing aPKC-Halo (no nuclear signal). Bottom: The cell cortex (F-actin) labelled with phalloidin-405. (C) Zoomed-in super-resolution images of the dashed boxes from (A). Left: a single membrane labelled since the juxtaposed membrane is located in a wild-type cell. Middle and right: both juxtaposed membranes labelled. (D) The principle behind the one labelled cortex and two labelled cortices using wild-type cell clones.

Starting with tagged aPKC-Halo, I analysed cell junctions where both juxtaposed membranes were labelled. Additionally, I also created wild-type cell clones in the tissue expressing aPKC-Halo and analysed the clone border. Along the clone border only a single membrane is labelled since one membrane comes from a cell expressing aPKC-Halo, which is competent for labelling, and the other cell expresses a non-tagged aPKC (Figure 5.8 and Figure 5.9). In both cases aPKC-Halo appeared non-homogeneously distributed (Figure 5.8C and Figure 5.9C) as previously observed with confocal microscopy (see Chapter 3).

Along the single membrane labelled junctions the aPKC-Halo molecules were distributed with an average density of 667 molecules per μm^2 (sd=141) (n=19). As expected the density almost doubled along the junctions where both juxtaposed membranes were labelled with an average density of 1094 molecules per μm^2 (sd=236) (n=77). The cytoplasmic density was on average 12 molecules per μm^2 (sd=8.12) (n=58) Figure 5.10A).

Next I wanted to analyse the cluster size distribution of aPKC-Halo molecules along the junction and approximate this distribution with mathematical functions as demonstrated on the simulated data before. Using the mean-shift algorithm with a bandwidth of 50 nm, the cluster size distribution extended from small clusters of a few molecules to a few clusters of hundreds of molecules (Figure 5.10B and Figure 5.10D). This distribution was best approximated with power-law function. This was the case for both single- and double-labelled membranes (Figure 5.10C and Figure 5.10E).

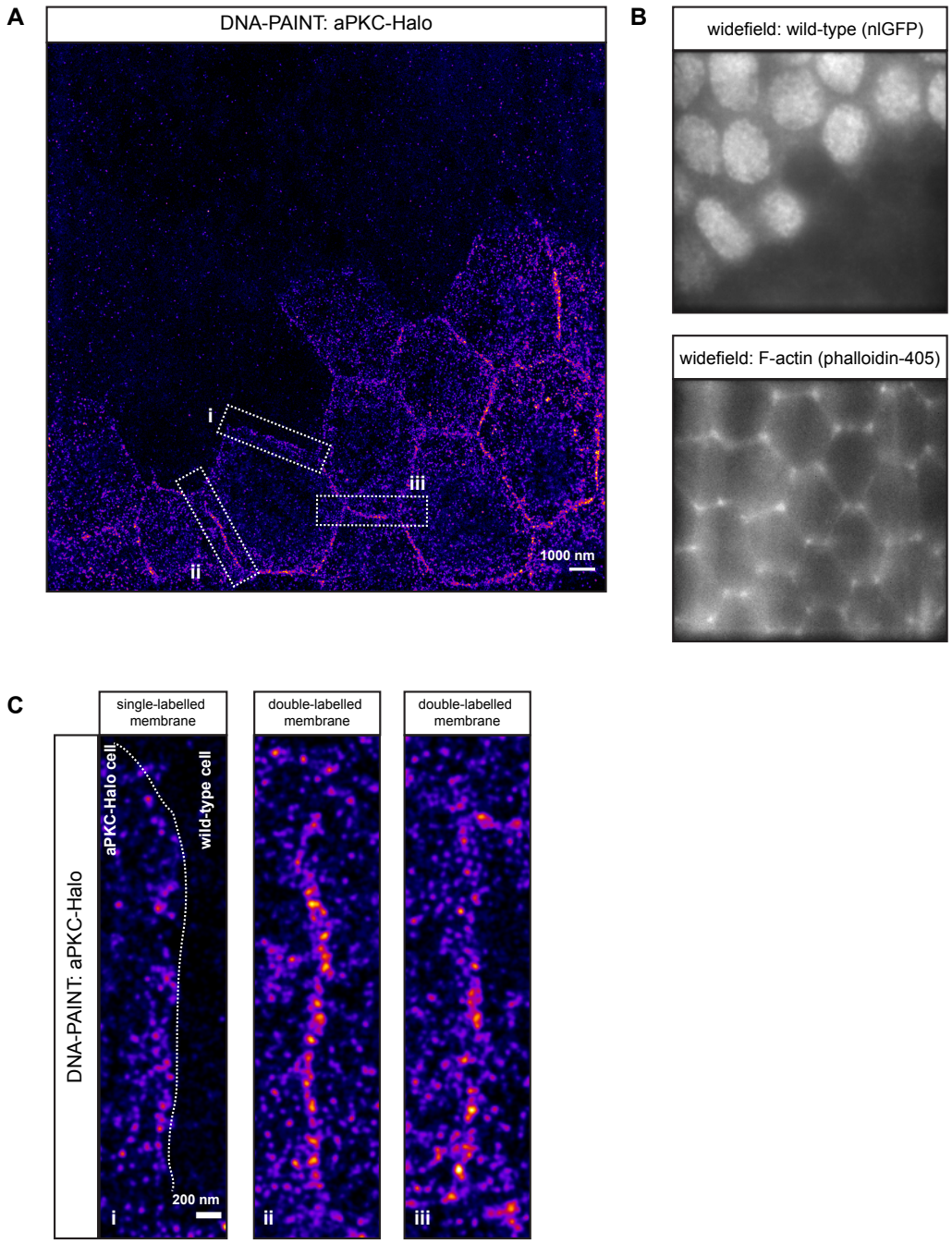


Figure 5.9 **A super-resolution image of aPKC-Halo using DNA-PAINT.** Legend on next page.

Figure 5.9 *(previous page)* (A) An example of a super-resolution image showing aPKC-Halo cells next to wild-type clone cells. (B) Top: the position of the wild-type cells (nuclear signal) relative to cell expressing aPKC-Halo (no nuclear signal). Bottom: the cell cortex (F-actin) labelled with phalloidin-405. (C) Zoomed-in super-resolution images of the dashed boxes from (A). Left: a single membrane labelled since the juxtaposed membrane is located in a wild-type cell. Middle and right: both juxtaposed membranes labelled.

Because the bandwidth was chosen arbitrary, I wondered how the cluster size distribution and hence the function approximation would change upon increasing it to 75nm and 100nm, respectively. Just visually inspecting the segmented junctions, I could observe that there are fewer smaller clusters upon using bigger bandwidth (Figure 5.11). This is expected since larger bandwidth results in smaller sensitivity for maxima of local densities (Comaniciu and Meer 2002). However, despite the shifted cumulative distribution function, the shape of the curve remained similar between all three bandwidths (Figure 5.11). Interestingly, with increasing the bandwidth, the cluster size distribution was now better approximated with the exponential function in single membrane labelled junctions (Figure 5.10C). The approximated function did not change in double membrane labelled junctions and remained power law despite an increase in bandwidth. However, the proportion of junctions in which the cluster size distribution was best approximated with the exponential function increased as well (Figure 5.10E).

I performed similar quantitative analysis on Crumbs- and Par6-Halo. However, because I had problems in creating wild-type cell clones I could only analyse junctions where both juxtaposed membranes were labelled. Like aPKC-Halo, both proteins appeared non-homogeneously distributed along the cell junctions (Figure 5.12).

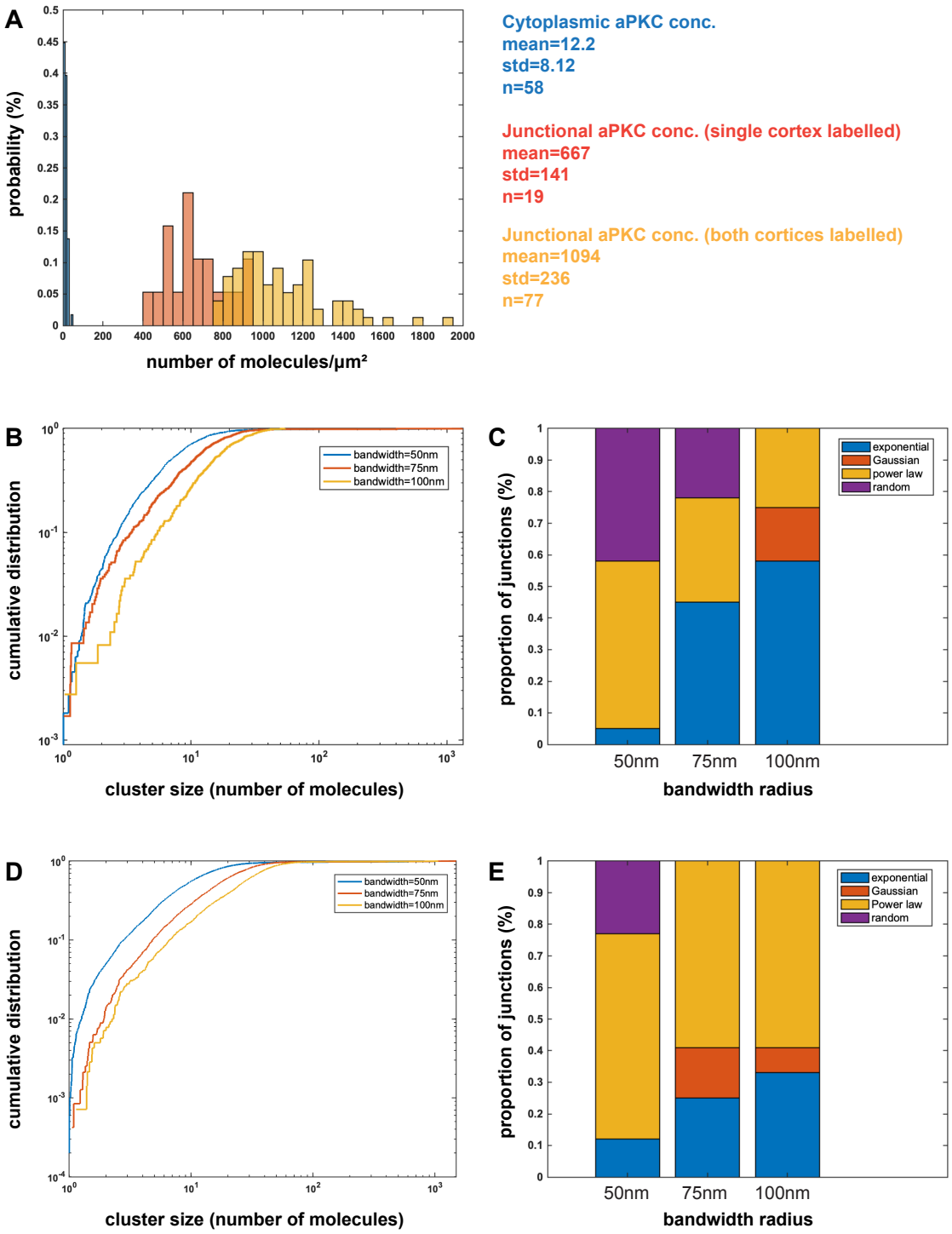


Figure 5.10 **Spatial analysis of aPKC-Halo in the follicle cells.** Legend on next page.

Figure 5.10 *(previous page)* (A) The distribution of the number of molecules per µm^2 in the cytoplasmic and junctional areas where a single or double membranes, respectively, were labelled. n denotes the number of analysed junctions in 3 independent experiments. (B) Cumulative distribution function of cluster sizes based on different bandwidths of the mean-shift algorithm from junctions where a single membrane was labelled. (C) Proportion of junctions, in which a single membrane was labelled, exhibiting different models of cluster size distribution based on different bandwidth of the mean-shift algorithm. (D) Cumulative distribution function of cluster sizes based on different bandwidth of the mean-shift algorithm from junctions where two juxtaposed membranes were labelled. (C) Proportion of junctions, where two juxtaposed membranes were labelled, exhibiting different models of cluster size distribution based on different bandwidths of the mean-shift algorithm.

For the density of Crumbs-Halo along the junctions where both juxtaposed membranes were labelled was on average of 533 molecules per µm^2 (sd=190) (n=25), while in the cytoplasm on average only 5 molecules were detected (sd=4.75) (n=42) (Figure 5.13A). Cluster size ranged from a few to a few hundreds of molecules (Figure 5.13B). The Crumbs-Halo cluster size distribution was well approximated by a power law (Figure 5.13C). With the bandwidth size of 75 nm the proportion of junctions where distribution was better approximated by exponential function increased. These values for Crumbs-Halo were calculated upon assuming 50% labelling efficiency, since the absolute labelling efficiency could not be experimentally determined (see Chapter 4).

For Par6-Halo the density of the junctions where both juxtaposed membranes were labelled was an average of 424 molecules per µm^2 (sd=136) (n=25), while in the cytoplasm an average of only 5 molecules were detected (sd=3.60) (n=43) (Figure 5.14A). Like Crumbs-Halo, the clusters were very polydispersed in size, ranging over three orders of magnitude. However, the majority of clusters contained fewer than 10 molecules (Figure 5.14B). The Par6-Halo cluster size distribution was well approximated by a power law, and this was not influenced by increasing the bandwidth size (Figure 5.14C).

Altogether these results suggest that all three investigated polarity proteins form clusters that are highly concentrated along the cell junctions (Table 5.1). These clusters may contain from a few up to thousands protein molecules. Their distribution is best approximated with a power-law function. While increasing the bandwidth in the mean-shift algorithm shifts the clusters towards bigger sizes, the model selection does not significantly change (Table 5.2).

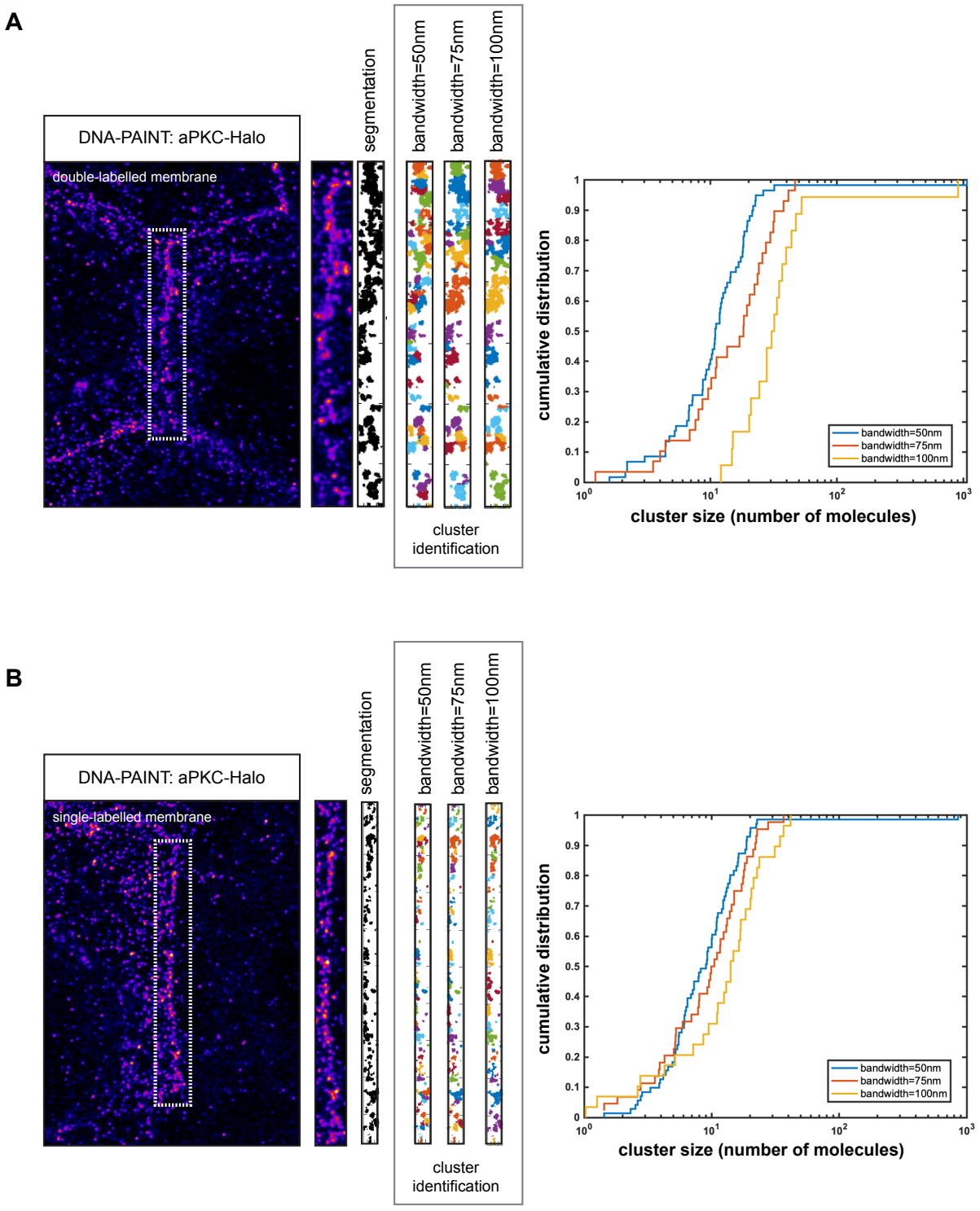


Figure 5.11 **Effect of bandwidth in the mean-shift algorithm on cluster size distribution.** Legend on next page.

Figure 5.11 *(previous page)* (A) A super-resolution image of aPKC-Halo at a junction where two juxtaposed membranes were labelled. For cluster segmentation only the dashed-boxed region is used. Cluster identification based on the bandwidth of the mean-shift algorithm Localisations belonging to the same cluster are colour-coded. The cumulative distribution function of cluster size based on different bandwidth of the mean-shift algorithm. (B) A super-resolution image of aPKC-Halo at the junction where a single membrane was labelled. For cluster segmentation only the dashed-boxed region is used. Cluster identification based on the bandwidth of the mean-shift algorithm. Localisations belonging to the same cluster are colour-coded. The cumulative distribution function of cluster size based on different bandwidths of the mean-shift algorithm.

5.4 Discussion

In this chapter I corroborated previous observations in confocal experiments (see Chapter 3) that apical polarity proteins are not homogeneously distributed along the cell junctions. The data presented here suggest that polarity proteins rather form clusters and that these vary in size in terms of number of proteins. I quantified the number of proteins per cluster and expanded this quantification to approximate the cluster size distribution with a mathematical function.

In spatial statistics, a big question is how clustering is defined still remains. The answer might be straightforward with well isolated cluster islands, but how one defines clusters that are less well separated, as is the case with polarity proteins, seems to be a bit more arbitrary. One should not forget that the human brain is especially apt in seeing spatial point patterns that might not be so easy to identify computationally. Multiple improvements to current approaches have been proposed, ranging from Bayesian cluster identification (Rubin-Delanchy et al. 2015) to machine learning (Williamson et al. 2018). A standard mean-shift algorithm was used in this study since it performed well and has been used in similar contexts before (Truong Quang et al. 2013).

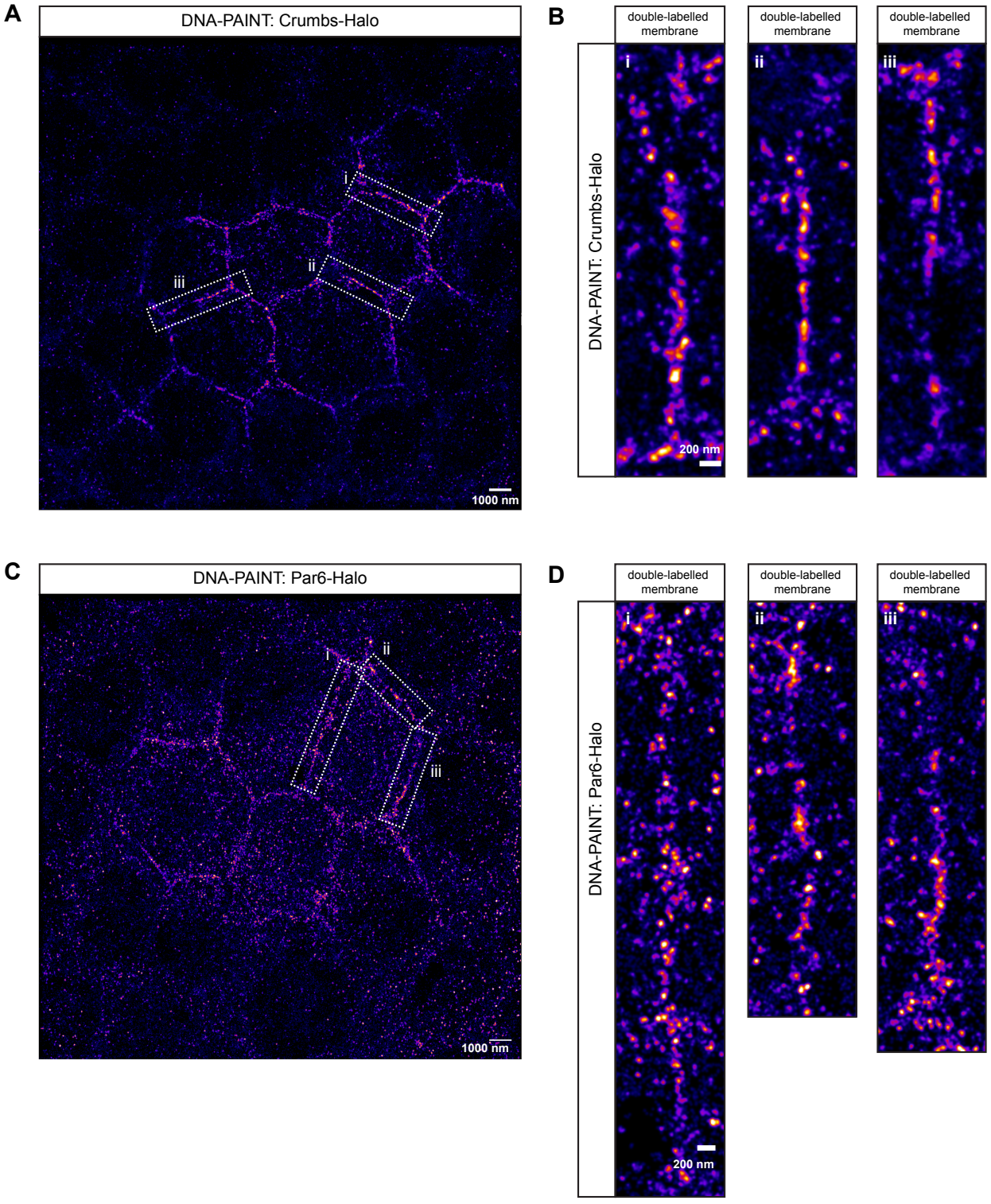


Figure 5.12 **A super-resolution image of Crumbs-Halo and Par-6 in the follicle cells using DNA-PAINT.** Legend on next page.

Figure 5.12 *(previous page)* (A) An example of a super-resolution image showing Crumbs-Halo in the marginal zone of the follicle cells. The dark part of the image is out of the focal plane because of tissue curvature. (B) Zoomed-in super-resolution image is of the dashed boxes from (A). (C) An example of a super-resolution image showing Par6-Halo in the marginal zone of the follicle cells. The dark part of the image are out of the focal plane because of tissue curvature. (D) Zoomed-in super-resolution images of the dashed boxes from (A).

Upon developing the analysis pipeline for selection of the model that would mathematically describe the cluster size distribution, the simulation parameters were chosen arbitrarily. It will be important to repeat the simulations with similar numbers of molecules per junctions as detected in the experiments, although the simulations partially covered this. For example, the simulated junctions had the area of 2.5 µm^2 and the number of simulated proteins per junctions was on average from 600 to 3000 (depending on the parameters). In the case of aPKC the average number of molecules per junction where a single membrane was labelled was around 1750 molecules per 2.5 µm^2.

Analysis of cluster size distributions requires stringent handling as measured distributions can be biased by image processing. For instance, thresholding (consider or eliminate clusters smaller than a certain size) can lead to contradictory conclusions on cluster size (Sherman et al. 2011). However, in our analyses we did not eliminate clusters smaller then a certain size but included them all. It will be important to investigate how the model selection would change if only clusters containing 10 or more proteins were considered.

The weak point in the study presented here is the lack of analysis of an experimental example of random cluster distribution. While I generated a transgenic line that expresses CAAX-Halo construct that should not oligomerise and be randomly distributed within the membrane, I did not acquire good quality images due to experimental difficulties. This experimental control for clustering should be done in the future. Moreover, imaging E-cadherin clusters could serve as a positive control. Its cluster size distribution follows a power-law (Truong Quang et al. 2013), which could be tested using our analysis approach.

Another weak point of this study is the analysis of double-labelled membranes. This could potentially lead to clusters being merged across the two membranes resulting in much bigger clusters in terms of the absolute number of molecules per cluster. This would be pronounced especially upon increased bandwidth sizes in the mean shift algorithm.

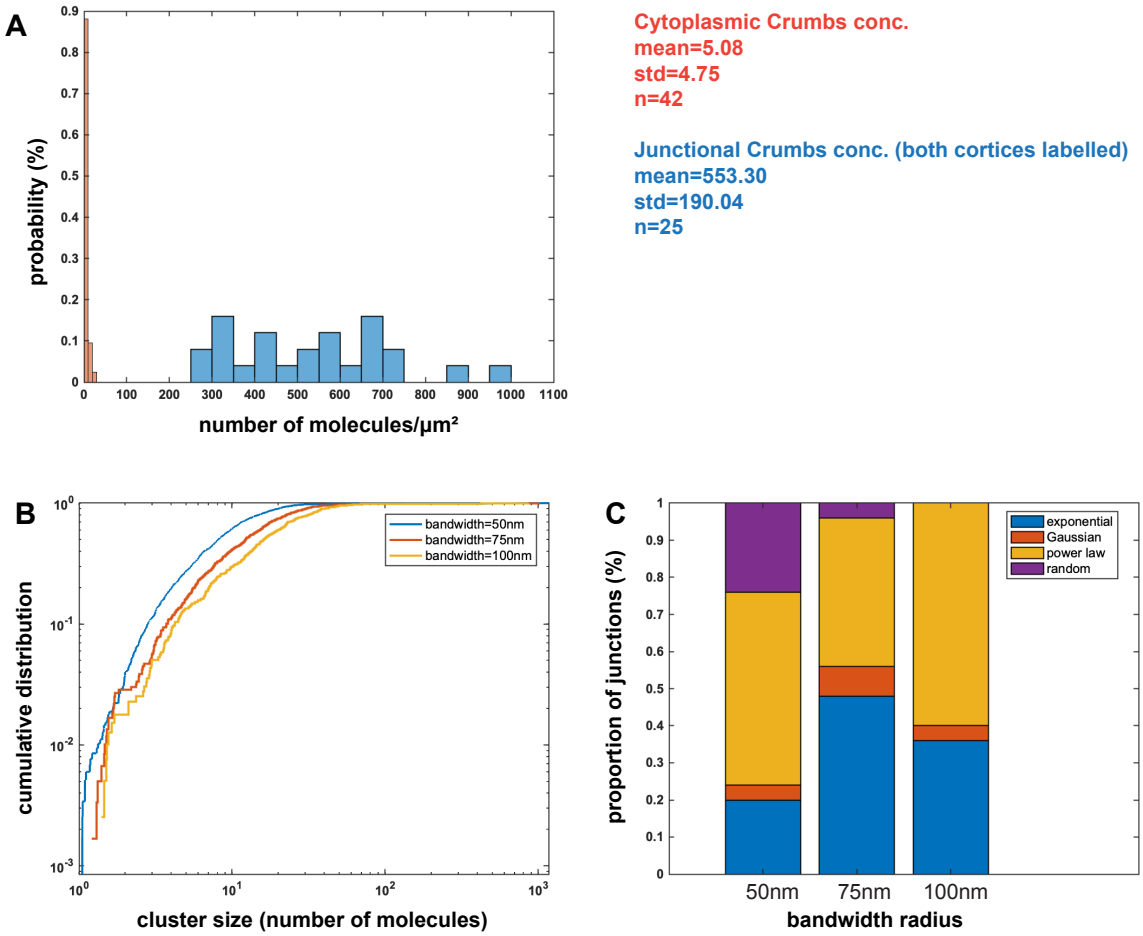


Figure 5.13 **Spatial analysis of Crumbs-Halo in the follicle cells.** Legend on next page.

However, based on the analysis of aPKC-Halo molecules in single- and double-labelled membranes this does not seem to affect the cluster size distribution.

Figure 5.13 *(previous page)* A) Distribution of number of Crumbs-Halo molecules per µm^2 in the cytoplasm and junctional areas where both juxtaposed membranes were labelled. n denotes the number of analysed junctions in 3 independent experiments. (B) Cumulative distribution function of cluster sizes based on the different bandwidths of the mean-shift algorithm from junctions where two juxtaposed membranes were labelled. (C) Proportion of junctions, where two juxtaposed membranes were labelled, exhibiting different models of cluster size distribution based on the different bandwidths of the mean-shift algorithm.

The majority of proteins I analysed exhibited cluster size distribution along the junctions that is well approximated by a power law, except for a few junctions exhibiting exponentially distributed cluster sizes. Importantly, the cluster size distribution was well approximated by a power law when all junctions were pooled together for each of the investigated polarity protein. How are power-law and exponential functions related? Speaking purely mathematically, in the power-law distribution the base is changing, while in the exponential distribution, the exponent is changing.

What can power-law and exponential function tell us about how these clusters occur? Both reflect dynamic fusion and fission processes. An exponential distribution of cluster size was observed in bacterial chemotaxis receptor clustering (Greenfield et al. 2009), with stochastic assembly by receptor diffusion as a mechanism. On the other hand, a power-law distribution of cluster size was observed in the organisation of E-cadherin molecules in Drosophila embryo (Truong Quang et al. 2013).

A power-law distribution has been implied in the diffusion-limited model of cluster-cluster aggregation (Meakin et al. 1985; Lin et al. 1989). This occurs when there is no repulsive force between the clusters and their fusion is limited solely by the time needed for clusters to encounter each other by diffusion. On the other hand, in the reaction-limited model, clustering occurs when there is a repulsive force, so that the clustering is limited by the time taken for two clusters to overcome this barrier (Lin et al. 1989).

A theoretical framework predicts that proteins of high valence, that is the ability to form multiple chemical bonds, form macrophases. Contrary, proteins of low valence, that is the ability to form small number of bonds, will form limited-size clusters. This is because proteins at the cluster edge cannot form new bonds (Markova et al. 2012).

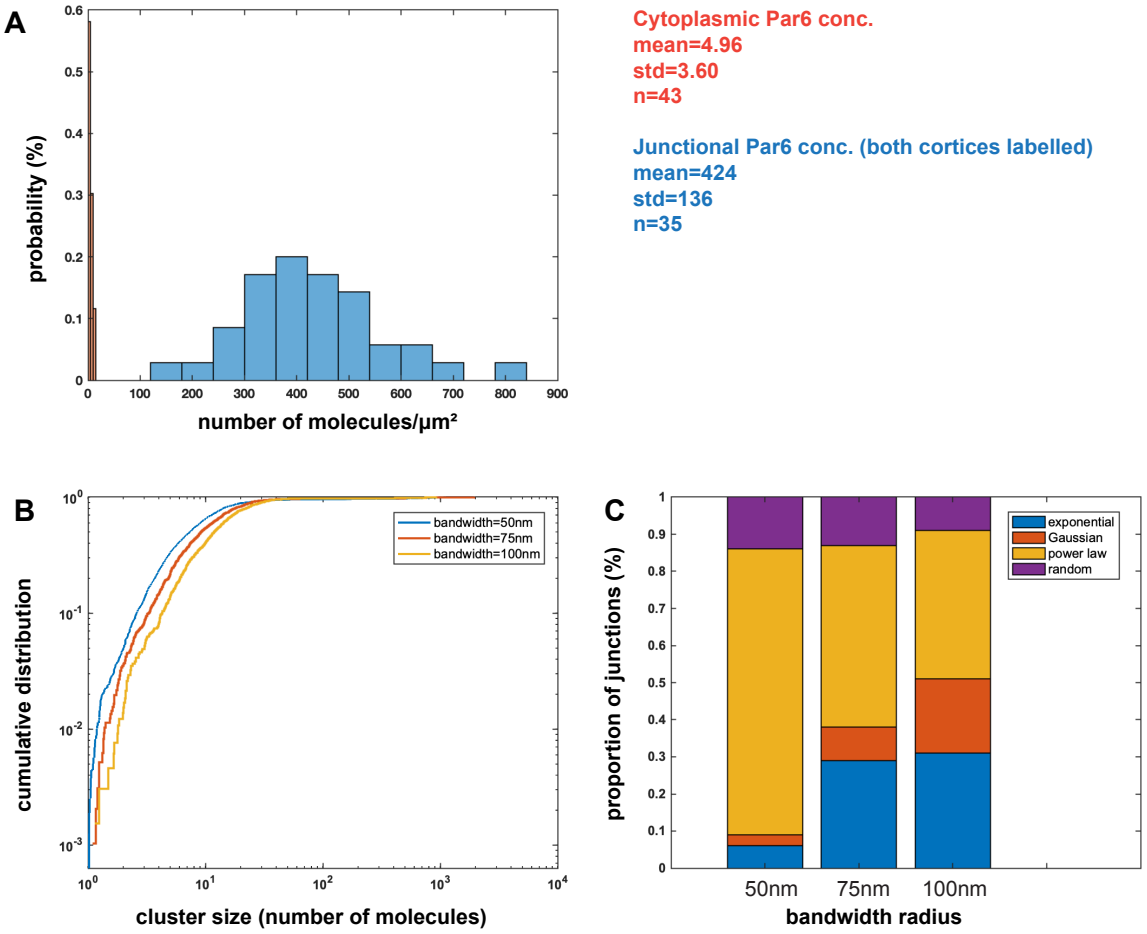


Figure 5.14 **Spatial analysis of Par6-Halo in the follicle cells.** Legend on next page.

Figure 5.14 *(previous page)* (A) Distribution of number of Par6-Halo molecules per µm^2 in the cytoplasm and junctional areas where both juxtaposed membranes were labelled. n denotes the number of analysed junctions in 3 independent experiments. (B) Cumulative distribution function of cluster sizes based on the different bandwidths of the mean-shift algorithm from junctions where two juxtaposed membranes were labelled. (C) Proportion of junctions, where two juxtaposed membranes were labelled, exhibiting different models of cluster size distribution based on the different bandwidths of the mean-shift algorithm.

Among the polarity proteins only two have been shown to self-oligomerise. Crumbs, through its the extracellular domain, and Par3 via its CR1 domain. How could clustering arise for aPKC and Par6? While all polarity proteins contain protein domains to form complexes, this can only explain hetero-dimers (aPKC-Par6) or hetero-oligomers (aPKC-Par6-Crumbs) but not higher order assemblies of these proteins. One explanation would be that these proteins form clusters through another unknown interacting protein. Phase separation is less likely, however this could be more thoroughly explored.

It will be important to validate the power law. This could be done by computational modelling of the clustering process where information about the molecule number, cluster number and diffusivity would be included. This would be especially informative about cluster maintenance and if a simple diffusion-limited model is enough. It is clear that clusters do not grow indefinitely, meaning that additional forces exist to restrict their sizes.

An important control experiment in the future would be to perform DNA-PAINT imaging using astigmatism. This would allow one to assess the three-dimensional shape of the clusters and determine if they are really spherical as I assumed. This is especially important since quantification of points within an three-dimensional object can be performed in two dimensions (planar sections) without the artefacts only if the object is spherical. This is also known as a Wicksell's corpuscle problem (Wicksell 1925). However, the investigated apical polarity proteins are enriched in the marginal zone, which is relatively narrow (less than a micrometer). Therefore it is valid to assume that the clusters probably do not form extremely heterogenous shapes, e.g. filaments.

Polarity protein	**Concentration (molecules / µm^2)**		
	Cytoplasmic	Junctional (single-labelled membrane)	Junctional (double-labelled membrane)
aPKC	12.2 ± 8.12 (n=58)	667 ± 141 (n=19)	1094 ± 236 (n=77)
Crumbs	5.08 ± 4.75 (n=42)	not assessed	553.30 ± 190.04 (n=25)
Par6	4.96 ± 3.60 (n=43)	not assessed	424 ± 136 (n=35)

Table 5.1: **A summary table of cytoplasmic and junctional concentrations for investigated polarity proteins.**

There is only one study that has addressed the copy number of polarity proteins in the context of polarity establishment, which examined the worm zygote. Using single-cell biochemistry, it was determined that aPKC is present in up to 15 copies per cluster during polarity establishment. The authors discussed that due to technical limitations this number is probably an underestimate (Dickinson et al. 2017). Interestingly, they established that 4.5% of total aPKC is present within the oligomers during polarity establishment, while only 1.5% of total aPKC is present within the oligomers during polarity maintenance.

In the case of the follicle epithelial cells the majority of clusters contained less than 20 molecules of proteins for all three investigated proteins. The interesting observation is that all junctions contained some clusters, which contained a few hundred proteins. In order to prove these observations are real one would need to perturb the system. In the case of Crumbs, which is a transmembrane protein, this could be possible by blocking its membrane recycling via endocytosis. There is a temperature-sensitive fly mutant for a dynamin, a GTPase that plays a role in clathrin-dependent endocytosis (Chen et al. 1991). Upon inactivating dynamin the endocytic vesicles are not pinched from the membrane and would accumulate over-time, which should be reflected in the cluster size distribution. A similar approach was used when analysing cluster sizes of E-cadherin in the Drosophila embryo (Truong Quang et al. 2013).

Moving from junctions to the cytoplasm, I quantified the cytoplasmic concentrations of apical polarity proteins within the marginal zone. There was no clustering detected in the cytoplasm and, interestingly, the density of the molecules was extremely low in comparison with that at junctions. It is tempting to suggest that the majority of investigated apical proteins are concentrated in the junctions in the form of clusters. This is different to what was observed in the worm zygote, where only a small proportion of anterior polarity proteins are within clusters (Dickinson et al. 2017). However, one could also argue that the fixation and permeabilisation treatment of the sample used in this work washed away a substantial amount of the cytoplasmic proteins, which would make the interpretation about the cytoplasmic stoichometry much more difficult.

It would be interesting to investigate how the cytoplasmic concentration changes along the apical-basal axis and if there is a sharp boundary between the apical and the lateral domain as it exist between polarity domains in worm zygote (Kay and Hunter 2001).

Polarity protein	**Bandwidth radius**	**Model selection for protein distribution (% of junctions)**			
		Exponential	Gaussian	Power-law	Random
aPKC (n=77)	50nm	12%	0%	65%	23%
aPKC (n=77)	75nm	25%	15%	60%	0%
aPKC (n=77)	100nm	33%	8%	59%	0%
Crumbs (n=25)	50nm	20%	4%	52%	24%
Crumbs (n=25)	75nm	48%	8%	40%	4%
Crumbs (n=25)	100nm	36%	4%	60%	0%
Par6 (n=35)	50nm	6%	3%	77%	14%
Par6 (n=35)	75nm	29%	9%	49%	14%
Par6 (n=35)	100nm	31%	20%	40%	11%

Table 5.2: **A summary table of model selection for protein distribution for three different bandwidth radii when using mean shift algorithm.**

Without a doubt, the computer simulations present in this work were simplified from the perspective of membrane organisation. The presence of the actin cortex was ignored, as long as the fact that the membranes are not flat but ruffled. In the future this two issues could be addressed since they can both help with understanding the origin of the clustered organisations.

For the actin cortex peptide-PAINT (LifeAct probe) can be used to qualitatively visualise F-actin (Kiuchi et al. 2015). Labelling the membrane is experimentally a bit more challenging. While there has been progress in developing of dyes for live cell labelling of membranes (Stone et al. 2017), there is still no good dye for fixed membranes. This problem could be circumvented by using nanobodies against CAAX-GFP that is incorporated into membranes (Kay and Hunter 2001)).

For example, in the case of E-cadherin it was observed that its clusters are actin-delimited in polarized Eph4 mammary epithelial cells (Wu 2015). However, this is probably not the case with investigate polarity proteins in the follicle cells since clusters are polydisperse in size. In case they would be actin-delimited the cluster distribution should be well approximated by Gaussian distribution, where the mean size would correspond to the size of the actin mesh.

Lastly, the results presented here suggest one more important point when it comes to understanding apical polarity complexes in Drosophila epithelial cells. The classical understanding is that aPKC-Par6 is complexed with Crumbs and that this is the main complex species (Fletcher et al. 2012). However, based on the quantifications presented here, it seems that there is at least twice as much aPKC as there is Crumbs and Par6, respectively. This means that even if the entire junctional fraction of Par6 is in a complex with aPKC, there is still plenty of aPKC left that is not bound to Par6.

For a long time it has been thought that aPKC directly interacts with Crumbs, however it has been recently showed that this putative phosphorylation does not occur (Cao et al. 2017). Moreover, aPKC has to be in complex with Par6 in order to be activated by pseudosubstrate displacement (Graybill et al. 2012). One possible explanation would be then that the remaining aPKC is simply in inactive state. The second explanation would be that this aPKC is coupled with Par3 and is waiting until activated by Par6 to phosphorylate Par3 as previously suggested (Morais-de-Sá et al. 2010).

It would be interesting to investigate if aPKC that is not in a complex with Par6 is monomeric, which would suggest a diffusive state, or in clusters, which would suggest a more immobile state. The latter would be similar to worm zygote where clustered aPKC is inactive (Rodriguez et al. 2017). This would also hint that there are heterogeneities in cluster species. To investigate how the inhibition of aPKC's kinase activity influences its spatial distribution and distribution of aPKC substrates an analogue-sensitive allele could be exploited (Hannaford et al. 2019).

The definite answers to all these burning questions about polarity complexes will be given by utilizing two-colour super-resolution imaging.

5.5 Perspectives

In conclusion, the findings in this chapter give more credit to the hypothesis that spatial organisation (i.e. clustering) of polarity proteins might be important not only in the worm zygote but also in other experimental systems. Moreover, polarity protein clustering in the worm zygote plays an important role during polarity establishment but it mainly disappears during its maintenance (except for CHIN1 protein). In fruit fly follicle cells clustering of all investigated polarity proteins is present during maintenance phase, which might suggest a functional role.

In terms of biological findings, I would like to highlight that the investigated apical polarity proteins (aPKC, Par6, Crumbs) seem to form clusters of polydisperse sizes. This distribution is well approximated by a power law function based on Bayesian-based model selection. This suggests that these clusters arise from diffusion-limited processes. Moreover, there seems to be considerably more aPKC molecules than Par6 and Crumbs, respectively, suggesting that not all aPKC is complexed with Par6.

Finally, this chapter can be also seen as a test tube for testing the post-processing processing approach to identify and mathematically describe molecular clustering (Figure 5.15). To my knowledge this is the first time that this kind of analysis of super-resolution images has been performed on junctional (non-membrane) proteins.

Altogether, this data will prove useful for modelling the spatial organisation of polarity proteins, and provide a framework for greater insight into the biological function of individual proteins.

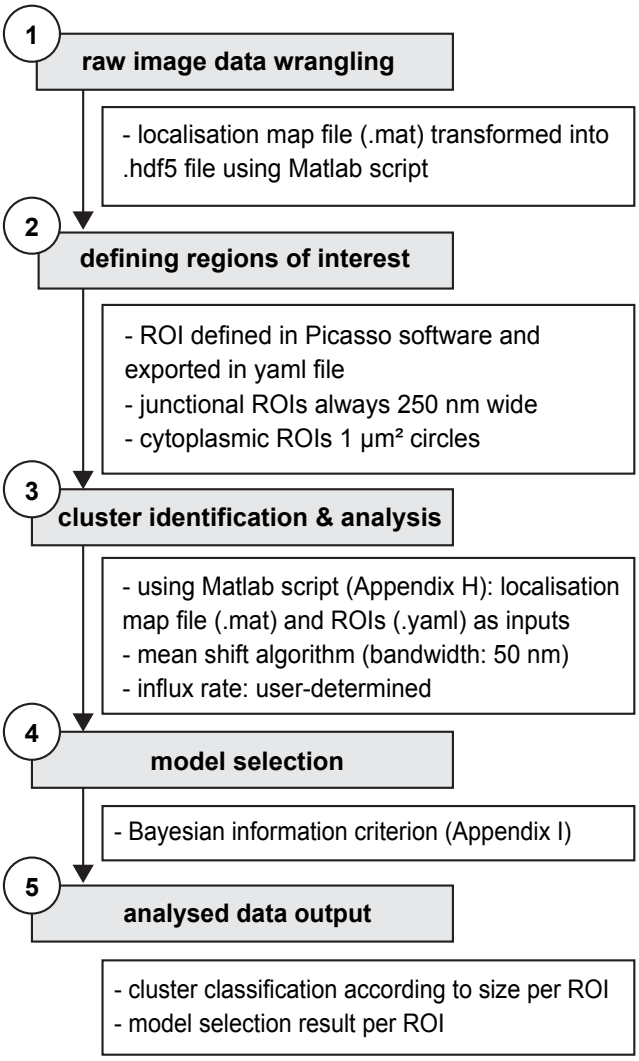


Figure 5.15 **Work-flow for the computer analysis of protein spatial organisation in super-resolution images.**

5.6 Acknowledgment of contributions

The computer simulations and the cluster analysis pipeline were developed and performed in collaboration with Leila Muresan, Cambridge Advanced Imaging Centre. The transgenic fly stocks expressing endogenously-tagged polarity proteins were created by Nick Lowe.

Chapter 6

Discussion

Utilising quantitative approaches to tackle biological problems has expanded the toolkit with which discoveries can be made. Not surprisingly, the imaging tools used in cell biology have mainly focused on standard confocal fluorescence microscopy.

However, with the arrival of super-resolution imaging approaches, we are not far from making imaging at the nano-scale as widespread as the current standard of imaging biological processes at the micro-scale. Whilst the early stages of the field of super-resolution only relied on the imaging of standard sample structures (e.g. microtubules, mitochondria), new biological insights have been also delivered more recently. These include the discovery of a trans-synaptic nanocolumn that aligns neurotransmitter release to receptors (Tang et al. 2016), the deciphering of the nanoscale architecture of cadherin-based adhesions (Bertocchi et al. 2017), and a description of the mechanisms governing assembly of the myosin filaments (Beach et al. 2017).

In this work I contribute to the mosaic of new biological insights using super-resolution techniques to describe the nanoscale organisation of apical polarity proteins in the follicular epithelium of fruit fly. Whilst polarity proteins have been previously investigated in a single-cell worm embryo, here I use DNA-PAINT for the first time to observe their spatial organisation in epithelial cells within the tissue.

Virtually all super-resolution imaging studies conducted so far have been using cultured cells as a sample. There are a few examples where dSTORM was used in tissue samples (Woodhams et al. 2019; Heller et al. 2017; Herrmannsdörfer et al. 2017; Hu et al. 2016; Hou et al. 2014) and only a couple with DNA-PAINT (Jungmann et al. 2016; Park et al. 2018).

In this thesis I firstly started investigating apical polarity proteins by using standard confocal fluorescence microscopy, because it is first important to define the biological question on a microscale. This is not only because super-resolution imaging is technically much more challenging, but also because some questions can be answered without necessity to increase the imaging resolution.

Using confocal imaging I observed that apical polarity proteins are not homogeneously distributed along the cell junctions. Performing fluorescence recovery after photobleaching (FRAP) experiments revealed that the clustered aPKC fluorescence signal is less mobile than the non-clustered aPKC signal. This was an important hint towards the potential functional significance of clustering. Many questions have been raised. Is clustering a way of keeping polarity proteins in place, so that they do not diffuse away? Or do they simply present a platform for bringing proteins together and facilitating further biochemical reactions?

Similar questions have been addressed in the field of planar cell polarity (polarisation of cells within the plane of a cell sheet). It was shown that core transmembrane planer cell polarity proteins are clustered and highly stable. This allows for cytoplasmic planar cell polarity proteins to be subsequently concentrated at transmembrane protein clusters (Strutt et al. 2011) with endocytosis modulating their levels (Cho et al. 2015) and phosphorylation their precise localisation (Strutt et al. 2019).

To start exploring clustering of apical polarity proteins I then first asked what is the size of these clusters. This is one of the most simple questions that can be addressed using super-resolution imaging, and can consequently also give a hint about their emergence. I first tested the dSTORM approach, however due to unsatisfactory preliminary results I then switched to using DNA-PAINT. I showed that DNA-PAINT is superior to dSTORM imaging in terms of immunity to photobleaching and quantification of target molecules.

In the original publication describing quantitative DNA-PAINT (Jungmann et al. 2016) counting molecules was performed by using the presumably single isolated target protein to calibrate the influx rate. In the case of NPCs, they labelled Nup98 using a monoclonal primary antibody directly conjugated to a docking oligo. Then they used a single isolated Nup98 protein clusters (or NPC subunits) as the calibration for the imager oligo influx rate. Here, I extracted this information by a mathematical approach of fitting the distribution

of multiple different mean values of t_{OFF} to extract the mean t_{OFF} for a single binding site.

At the time of submission of this thesis a new technique suggesting counting of the absolute protein numbers emerged. Localisation-based fluorescence correlation spectroscopy (lbFCS) is an approach where extracting imager oligo hybridisation rates is independent from knowing the number of binding sites (Stein et al. 2019). This approach works by analysing the binding kinetics of well-separated clusters at different concentrations of imager strands. However, this approach has been tested on DNA-origami so far, using low laser powers and low numbers of binding sites in the clusters. If this can be used in a biological sample with high background fluorescence and high number of binding sites per cluster is yet to be determined.

I believe that the NPC presents an in vivo "origami" structure and is, to date, the best structure to calibrate the influx rate for DNA-PAINT quantifications. One way to circumvent the possible errors due to the fitting approach would be to perform the binding kinetics on Gle1 protein, which is a protein in the NPC subunits of the cytoplasmic ring (Ori et al. 2013). Importantly, only one copy is present per subunit. Having one binding site per subunit means that one would just need to analyse the binding kinetics and determine the mean t_{OFF} of the subunit area. Gle1 is also a relatively small protein (77 kDa), which would enable assaying ALE with the gel band shift assay.

A second novelty presented here that adds to the general method of quantitative DNA-PAINT is determining the absolute labelling efficiency using a gel band shift assay. Since the molecular target population is usually not fully labelled, this information is crucial to quantify the absolute number of proteins without undercounting.

Using the above counting approach I then quantified the absolute number of proteins within the clusters. It seems clear that the cluster size of investigated proteins cannot be described with an average number of molecules per cluster, since it spans over three magnitudes of size. However, whether cluster size distribution really follows a power-law function has yet to be further validated. More junctions would need to be analysed, and especially with a single-labelled membrane.

An important next experimental step is two-colour imaging. In this thesis I showed that two-colour DNA-PAINT imaging is possible on the example of the nuclear pore complex.

I performed some preliminary two-colour imaging on apical polarity proteins as well, however I decided to not include this data in my thesis, specifically because of the two reasons. Firstly, because of the low and different labelling efficiencies for the proteins of interest co-localisation analysis is almost impossible. This is the case because the absence of the fluorescent signal co-localisation does not necessarily mean that the two proteins do not co-localise but it might be that one of them is just not labelled. Secondly, two-colour image registration is a troublesome process that at the time of performing two-colour imaging was not yet optimised for the imaging system used in this thesis. This meant that co-localisation was not perfect and it would be difficult to objectively analyse if the absence of co-localisation is of the biological or the optical origin. Nevertheless, for the future experiments, when two-colour imaging will be technically more feasible I suggest to investigate the correlation between the cluster size distribution of aPKC-Crumbs and Par6-Crumbs clusters. Since Crumbs is proposed to oligomerise it is logical that its cytoplasmic interactors (aPKC-Par6 complex) will also recapitulate the cluster size distribution of Crumbs molecules.

However, since aPKC is present in a much higher concentration along the junctions, it would be important to correlate whether aPKC that does not co-localise with Crumbs might follow a different cluster size distribution, assuming that it still clusters. An important protein to image as well is Par3. This is not only because it can oligomerise and would serve as a positive control, but also to correlate its spatial distribution with aPKC.

How could we explore the dynamics of cluster emergence process in a fixed tissue? I suggest investigating mitotic cells. It has been observed that in the follicular epithelium mitotic cells round up and polarity proteins de-localise into the cytoplasm and re-localise at the end of cell division (Morais-De-Sá and Sunkel 2013; Carvalho et al. 2015; Bergstralh et al. 2013). The fluorescence intensity of the labelled proteins also drops during mitosis, suggesting that they might also de-cluster. If this is the case, then imaging mitotic cells at different time points (based on chromatin shape one can classify mitotic phase) could reveal how clusters re-emerge after cell division. Moreover it would be interesting to investigate other epithelial tissues in fruit fly. Embryonic epithelia seems attractive to start with since it polarises during cellularisation and the apical side faces the cuticle that surrounds the membrane. In terms of imaging this is favourable since the target proteins do not reside deep in the tissue as in the follicular epithelium.

I hope that the imaging pipeline and the post-processing analysis presented in this thesis

will not only help to drive further research on polarity protein spatial organisation, but also in other areas of cell biology in thick samples. A few potential examples are performing quantitative super-resolution imaging using DNA-PAINT to investigate the stoichiometry of proteins which build non-centrosomal microtubule organising centres (Nashchekin et al. 2016) or experimentally validating the modelling approaches of the maximum cargo capacity of intracellular transport vesicles (Martins Ratamero and Royle 2019).

Bibliography

Abbe, E. (1873). "Beiträge zur Theorie des Mikroskops und der mikroskopischen Wahrnehmung". In: *Arch. für Mikroskopische Anat.*

Abney, J. R., J. Braun, and J. C. Owicki (1987). "Lateral interactions among membrane proteins. Implications for the organization of gap junctions". In: *Biophys. J.*

Achilleos, A., A. M. Wehman, and J. Nance (June 2010). "PAR-3 mediates the initial clustering and apical localization of junction and polarity proteins during C. elegans intestinal epithelial cell polarization." In: *Development* 137.11, pp. 1833–1842.

Acuna, G. P., F. M. Möller, P. Holzmeister, S. Beater, B. Lalkens, and P. Tinnefeld (2012). "Fluorescence enhancement at docking sites of DNA-directed self-assembled nanoantennas". In: *Science (80-.).*

Adams, A. E., D. I. Johnson, R. M. Longnecker, B. F. Sloat, and J. R. Pringle (1990). "CDC42 and CDC43, two additional genes involved in budding and the establishment of cell polarity in the yeast Saccharomyces cerevisiae". In: *J. Cell Biol.*

Alber, F. et al. (2007). "The molecular architecture of the nuclear pore complex." In: *Nature.*

Arata, Y. et al. (Aug. 2016). "Cortical Polarity of the RING Protein PAR-2 Is Maintained by Exchange Rate Kinetics at the Cortical-Cytoplasmic Boundary." In: *Cell Rep.* 16.8, pp. 2156–2168.

Auer, A., M. T. Strauss, T. Schlichthaerle, and R. Jungmann (2017). "Fast, Background-Free DNA-PAINT Imaging Using FRET-Based Probes". In: *Nano Lett.*

Banani, S. F., H. O. Lee, A. A. Hyman, and M. K. Rosen (Feb. 2017). "Biomolecular condensates: organizers of cellular biochemistry." In: *Nat. Rev. Mol. Cell Biol.* 18.5, pp. 285–298.

Banjade, S. and M. K. Rosen (Oct. 2014). "Phase transitions of multivalent proteins can promote clustering of membrane receptors." In: *Elife* 3, p. 641.

Bassett, A. R., C. Tibbit, C. P. Ponting, and J. L. Liu (2013). "Highly Efficient Targeted Mutagenesis of Drosophila with the CRISPR/Cas9 System". In: *Cell Rep.*

Bates, M., T. R. Blosser, and X. Zhuang (2005). "Short-range spectroscopic ruler based on a single-molecule optical switch". In: *Phys. Rev. Lett.*

Baumgart, F., A. M. Arnold, B. K. Rossboth, M. Brameshuber, and G. J. Schütz (Nov. 2018). "What we talk about when we talk about nanoclusters." In: *Methods Appl. Fluoresc.* 7.1, p. 13001.

Bayraktar, J., D. Zygmunt, and R. W. Carthew (2006). "Par-1 kinase establishes cell polarity and functions in Notch signaling in the Drosophila embryo". In: *J. Cell Sci.*

Beach, J. R. et al. (2017). "Actin dynamics and competition for myosin monomer govern the sequential amplification of myosin filaments". In: *Nat. Cell Biol.*

Benton, R. and D. St Johnston (Aug. 2003). "A conserved oligomerization domain in drosophila Bazooka/PAR-3 is important for apical localization and epithelial polarity." In: *Curr. Biol.* 13.15, pp. 1330–1334.

Bergstralh, D. T., T. Haack, and D. St Johnston (Jan. 2013). "Epithelial polarity and spindle orientation: intersecting pathways." In: *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 368.1629, p. 20130291.

Bernardo, J. M. and A. F. Smith (2008). *Bayesian Theory.*

Bertocchi, C. et al. (2017). "Nanoscale architecture of cadherin-based cell adhesions". In: *Nat. Cell Biol.*

Betzig, E., G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J. S. Bonifacino, M. W. Davidson, J. Lippincott-Schwartz, and H. F. Hess (2006). "Imaging intracellular fluorescent proteins at nanometer resolution". In: *Science (80-.).*

Bienkowska, D. and C. R. Cowan (2012). "Centrosomes can initiate a polarity axis from any position within one-cell C. Elegans embryos". In: *Curr. Biol.*

Bilder, D., M. Li, and N. Perrimon (2000). "Cooperative regulation of cell polarity and growth by Drosophila tumor suppressors". In: *Science (80-.).*

Blumhardt, P., J. Stein, J. Mücksch, F. Stehr, J. Bauer, R. Jungmann, and P. Schwille (Nov. 2018). "Photo-Induced Depletion of Binding Sites in DNA-PAINT Microscopy." In: *Molecules* 23.12, p. 3165.

Boyd, L., S. Quo, D. Levitan, D. T. Stinchcomb, and K. J. Kemphues (1996). "PAR-2 is asymmetrically distributed and promotes association of P granules and PAR-1 with the cortex in C. elegans embryos". In: *Development.*

Braun, J., J. R. Abney, and J. C. Owicki (1987). “Lateral interactions among membrane proteins. Valid estimates based on freeze-fracture electron microscopy”. In: *Biophys. J.*

Buszczak, M. et al. (2007). “The carnegie protein trap library: A versatile tool for drosophila developmental studies”. In: *Genetics*.

Campbell, K., E. Knust, and H. Skaer (2009). “Crumbs stabilises epithelial polarity during tissue remodelling”. In: *J. Cell Sci.*

Cao, H., R. Xu, Q. Shi, D. Zhang, J. Huang, and Y. Hong (Aug. 2017). “FERM domain phosphorylation and endogenous 3’UTR are not essential for regulating the function and subcellular localization of polarity protein Crumbs.” In: *J. Genet. Genomics* 44.8, pp. 409–412.

Carvalho, C. A., S. Moreira, G. Ventura, C. E. Sunkel, and E. Morais-De-Sá (2015). “Aurora a triggers Lgl cortical release during symmetric division to control planar spindle orientation”. In: *Curr. Biol.*

Chen, J., A. C. Sayadian, N. Lowe, H. E. Lovegrove, and D. St Johnston (2018). “An alternative mode of epithelial polarity in the Drosophila midgut”. In: *PLoS Biol.*

Chen, M. S., R. A. Obar, C. C. Schroeder, T. W. Austin, C. A. Poodry, S. C. Wadsworth, and R. B. Vallee (1991). “Multiple forms of dynamin are encoded by shibire, a Drosophila gene involved in endocytosis”. In: *Nature*.

Cho, B., G. Pierre-Louis, A. Sagner, S. Eaton, and J. D. Axelrod (2015). “Clustering and Negative Feedback by Endocytosis in Planar Cell Polarity Signaling Is Modulated by Ubiquitinylation of Prickle”. In: *PLoS Genet.*

Cisse, I. I., I. Izeddin, S. Z. Causse, L. Boudarene, A. Senecal, L. Muresan, C. Dugast-Darzacq, B. Hajj, M. Dahan, and X. Darzacq (Aug. 2013). “Real-time dynamics of RNA polymerase II clustering in live human cells.” In: *Science* 341.6146, pp. 664–667.

Comaniciu, D. and P. Meer (2002). “Mean shift: A robust approach toward feature space analysis”. In: *IEEE Trans. Pattern Anal. Mach. Intell.*

Cotteret, S. and J. Chernoff (2002). *The evolutionary history of effectors downstream of Cdc42 and Rac*.

Cowan, C. R. and A. A. Hyman (2004). “ ASYMMETRIC CELL DIVISION IN C. ELEGANS : Cortical Polarity and Spindle Positioning ”. In: *Annu. Rev. Cell Dev. Biol.*

Dawes, A. T. and E. M. Munro (Sept. 2011). “PAR-3 oligomerization may provide an actin-independent mechanism to maintain distinct par protein domains in the early Caenorhabditis elegans embryo.” In: *Biophys. J.* 101.6, pp. 1412–1422.

Dempsey, G. T., J. C. Vaughan, K. H. Chen, M. Bates, and X. Zhuang (Nov. 2011). "Evaluation of fluorophores for optimal performance in localization-based super-resolution imaging." In: *Nat. Methods* 8.12, pp. 1027–1036.

Deng, W. M., C. Althauser, and H. Ruohola-Baker (2001). "Notch-Delta signaling induces a transition from mitotic cell cycle to endocycle in Drosophila follicle cells". In: *Development*.

Desai, A. and T. J. Mitchison (1997). "MICROTUBULE POLYMERIZATION DYNAMICS". In: *Annu. Rev. Cell Dev. Biol.*

Deußner-Helfmann, N. S., A. Auer, M. T. Strauss, S. Malkusch, M. S. Dietz, H. D. Barth, R. Jungmann, and M. Heilemann (2018). "Correlative Single-Molecule FRET and DNA-PAINT Imaging". In: *Nano Lett.*

Dickinson, D. J., F. Schwager, L. Pintard, M. Gotta, and B. Goldstein (Aug. 2017). "A Single-Cell Biochemistry Approach Reveals PAR Complex Dynamics during Cell Polarization". In: *Dev. Cell* 42.4, 416–434.e11.

Dong, W., J. Lu, X. Zhang, Y. Wu, K. Lettieri, G. R. Hammond, and Y. Hong (Mar. 2019). "A Polybasic Domain in aPKC Mediates Par-6-Dependent Control of Plasma Membrane Targeting and Kinase Activity". In: *bioRxiv* 22, p. 588624.

Duffy, J. B. (2002). *GAL4 system in Drosophila: A fly geneticist's Swiss army knife*.

Dustin, M. L. and J. T. Groves (2012). "Receptor Signaling Clusters in the Immune Synapse". In: *Annu. Rev. Biophys.*

Erdmann, R. S. et al. (2019). "Labeling Strategies Matter for Super-Resolution Microscopy: A Comparison between HaloTags and SNAP-tags". In: *Cell Chem. Biol.*

Ester, M., H.-P. Kriegel, J. Sander, and X. Xu (1996). "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: *Proc. 2nd Int. Conf. Knowl. Discov. Data Min.*

Etemad-Moghadam, B., S. Guo, and K. J. Kemphues (1995). "Asymmetrically distributed PAR-3 protein contributes to cell polarity and spindle alignment in early C. elegans embryos". In: *Cell*.

Fletcher, G. C., E. P. Lucas, R. Brain, A. Tournier, and B. J. Thompson (June 2012). "Positive feedback and mutual antagonism combine to polarize Crumbs in the Drosophila follicle cell epithelium." In: *Curr. Biol.* 22.12, pp. 1116–1122.

Franz, A. and V. Riechmann (Feb. 2010). "Stepwise polarisation of the Drosophila follicular epithelium." In: *Dev. Biol.* 338.2, pp. 136–147.

Fukunaga, K. and L. D. Hostetler (1975). "The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition". In: *IEEE Trans. Inf. Theory*.

Gamblin, C. L., F. Parent-Prévost, K. Jacquet, C. Biehler, A. Jetté, and P. Laprise (2018). "Oligomerization of the FERM-FA protein Yurt controls epithelial cell polarity". In: *J. Cell Biol.*

Gamblin, C. L., É. J.-L. Hardy, F. J.-M. Chartier, N. Bisson, and P. Laprise (Feb. 2014). "A bidirectional antagonism between aPKC and Yurt regulates epithelial cell polarity." In: *J. Cell Biol.* 204.4, pp. 487–495.

Garcia-Parajo, M. F., A. Cambi, J. A. Torreno-Pina, N. Thompson, and K. Jacobson (Dec. 2014). "Nanoclustering as a dominant feature of plasma membrane organization." In: *J. Cell Sci.* 127.Pt 23, pp. 4995–5005.

Gavagnin, E., J. P. Owen, and C. A. Yates (2018). "Pair correlation functions for identifying spatial correlation in discrete domains". In: *Phys. Rev. E*.

Genova, J. L., S. Jong, J. T. Camp, and R. G. Fehon (May 2000). "Functional analysis of Cdc42 in actin filament assembly, epithelial morphogenesis, and cell signaling during Drosophila development." In: *Dev. Biol.* 221.1, pp. 181–194.

Goehring, N. W. and S. W. Grill (2013). *Cell polarity: Mechanochemical patterning*.

Goehring, N. W., P. K. Trong, J. S. Bois, D. Chowdhury, E. M. Nicola, A. A. Hyman, and S. W. Grill (2011a). "Polarization of PAR proteins by advective triggering of a pattern-forming system". In: *Science (80-.).*

Goehring, N. W., D. Chowdhury, A. A. Hyman, and S. W. Grill (Oct. 2010). "FRAP Analysis of Membrane-Associated Proteins: Lateral Diffusion and Membrane-Cytoplasmic Exchange". In: *Biophys. J.* 99.8, pp. 2443–2452.

Goehring, N. W., C. Hoege, S. W. Grill, and A. A. Hyman (May 2011b). "PAR proteins diffuse freely across the anterior-posterior boundary in polarized C. elegans embryos." In: *J. Cell Biol.* 193.3, pp. 583–594.

Gomez-Lamarca, M. J. et al. (Mar. 2018). "Activation of the Notch Signaling Pathway In Vivo Elicits Changes in CSL Nuclear Dynamics". In: *Dev. Cell* 44.5, 611–623.e7.

Graybill, C., B. Wee, S. X. Atwood, and K. E. Prehoda (June 2012). "Partitioning-defective protein 6 (Par-6) activates atypical protein kinase C (aPKC) by pseudosubstrate displacement." In: *J. Biol. Chem.* 287.25, pp. 21003–21011.

Greenfield, D., A. L. McEvoy, H. Shroff, G. E. Crooks, N. S. Wingreen, E. Betzig, and J. Liphardt (June 2009). "Self-organization of the Escherichia coli chemotaxis network

imaged with super-resolution light microscopy." In: *PLoS Biol.* 7.6. Ed. by H. C. Berg, e1000137.

Grossfield, A., S. E. Feller, and M. C. Pitman (2007). "Convergence of molecular dynamics simulations of membrane proteins". In: *Proteins Struct. Funct. Genet.*

Hannaford, M., N. Loyer, F. Tonelli, M. Zoltner, and J. Januschke (2019). "A chemical-genetics approach to study the role of atypical protein kinase C in Drosophila". In: *Dev.*

Hao, Y., Q. Du, X. Chen, Z. Zheng, J. L. Balsbaugh, S. Maitra, J. Shabanowitz, D. F. Hunt, and I. G. MacAra (2010). "Par3 controls epithelial spindle orientation by aPKC-mediated phosphorylation of apical pins". In: *Curr. Biol.*

Harris, T. J. C. and M. Peifer (Aug. 2005). "The positioning and segregation of apical cues during epithelial polarity establishment in Drosophila." In: *J. Cell Biol.* 170.5, pp. 813–823.

— (2007). "aPKC Controls Microtubule Organization to Balance Adherens Junction Symmetry and Planar Polarity during Development". In: *Dev. Cell.*

Harris, T. J. C. and U. Tepass (2010). *Adherens junctions: From molecules to morphogenesis.*

Hartman, N. C. and J. T. Groves (2011). *Signaling clusters in the cell membrane.*

Heilemann, M., S. Van De Linde, M. Schüttpelz, R. Kasper, B. Seefeldt, A. Mukherjee, P. Tinnefeld, and M. Sauer (2008). "Subdiffraction-resolution fluorescence imaging with conventional fluorescent probes". In: *Angew. Chemie - Int. Ed.*

Heller, J. P., P. Michaluk, K. Sugao, and D. A. Rusakov (2017). "Probing nano-organization of astroglia with multi-color super-resolution microscopy". In: *J. Neurosci. Res.*

Herrmannsdörfer, F., B. Flottmann, S. Nanguneri, V. Venkataramani, H. Horstmann, T. Kuner, and M. Heilemann (2017). "3D d STORM imaging of fixed brain tissue". In: *Methods Mol. Biol.*

Hess, S. T., T. P. Girirajan, and M. D. Mason (2006). "Ultra-high resolution imaging by fluorescence photoactivation localization microscopy". In: *Biophys. J.*

Hirai, T. and K. Chida (2003). *Protein kinase Cζ (PKCζ): Activation mechanisms and cellular functions.*

Holly, R. W. and K. E. Prehoda (2019). *Phosphorylation of Par-3 by Atypical Protein Kinase C and Competition between Its Substrates.*

Hong, Y., B. Stronach, N. Perrimon, L. Y. Jan, and Y. N. Jan (Dec. 2001). "Drosophila Stardust interacts with Crumbs to control polarity of epithelia but not neuroblasts". In: *Nature* 414.6864, pp. 634–638.

Horikoshi, Y., A. Suzuki, T. Yamanaka, K. Sasaki, K. Mizuno, H. Sawada, S. Yonemura, and S. Ohno (2009). "Interaction between PAR-3 and the aPKC-PAR-6 complex is indispensable for apical domain development of epithelial cells". In: *J. Cell Sci.*

Hou, Y., D. J. Crossman, V. Rajagopal, D. Baddeley, I. Jayasinghe, and C. Soeller (2014). "Super-resolution fluorescence imaging to study cardiac biophysics: α-actinin distribution and Z-disk topologies in optically thick cardiac tissue slices". In: *Prog. Biophys. Mol. Biol.*

Hu, Y. S., H. Cang, and B. F. Lillemeier (June 2016). "Superresolution imaging reveals nanometer- and micrometer-scale spatial distributions of T-cell receptors in lymph nodes." In: *Proc. Natl. Acad. Sci. U. S. A.* 113.26, pp. 7201–7206.

Huang, F. et al. (2013). "Video-rate nanoscopy using sCMOS camera-specific single-molecule localization algorithms". In: *Nat. Methods*.

Huang, J., W. Zhou, W. Dong, A. M. Watson, and Y. Hong (May 2009). "Directed, efficient, and versatile modifications of the Drosophila genome by genomic engineering." In: *Proc. Natl. Acad. Sci. U. S. A.* 106.20, pp. 8284–8289.

Huff, J. (2015). "The Airyscan detector from ZEISS: confocal imaging with improved signal-to-noise ratio and super-resolution". In: *Nat. Methods*.

Hutterer, A., J. Betschinger, M. Petronczki, and J. A. Knoblich (June 2004). "Sequential roles of Cdc42, Par-6, aPKC, and Lgl in the establishment of epithelial polarity during Drosophila embryogenesis." In: *Dev. Cell* 6.6, pp. 845–854.

Huynh, J. R., M. Petronczki, J. A. Knoblich, and D. S. Johnston (2001). "Bazooka and PAR-6 are required with PAR-1 for the maintenance of oocyte fate in Drosophila". In: *Curr. Biol.*

Hwang, M. S., J. Boulanger, J. D. Howe, A. Albecka, M. Pasche, L. Mureşan, and Y. Modis (2019). "MAVS polymers smaller than 80 nm induce mitochondrial membrane remodeling and interferon signaling". In: *FEBS J.*

Illian, J., A. Penttinen, H. Stoyan, and D. Stoyan (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*.

Jayasinghe, I., A. H. Clowsley, R. Lin, T. Lutz, C. Harrison, E. Green, D. Baddeley, L. Di Michele, and C. Soeller (Jan. 2018). "True Molecular Scale Visualization of Variable Clustering Properties of Ryanodine Receptors." In: *Cell Rep.* 22.2, pp. 557–567.

Jia, D., Q. Xu, Q. Xie, W. Mio, and W. M. Deng (2016). "Automatic stage identification of Drosophila egg chamber based on DAPI images". In: *Sci. Rep.*

Joberty, G., C. Petersen, L. Gao, and I. G. Macara (2000). "The cell-polarity protein Par6 links Par3 and atypical protein kinase C to Cdc42". In: *Nat. Cell Biol.*

Jungmann, R., M. S. Avendaño, J. B. Woehrstein, M. Dai, W. M. Shih, and P. Yin (Mar. 2014). "Multiplexed 3D cellular super-resolution imaging with DNA-PAINT and Exchange-PAINT". In: *Nat. Methods* 11.3, pp. 313–318.

Jungmann, R., M. S. Avendaño, M. Dai, J. B. Woehrstein, S. S. Agasti, Z. Feiger, A. Rodal, and P. Yin (Mar. 2016). "Quantitative super-resolution imaging with qPAINT". In: *Nat. Methods* 13.5, pp. 439–442.

Jungmann, R., C. Steinhauer, M. Scheible, A. Kuzyk, P. Tinnefeld, and F. C. Simmel (Nov. 2010). "Single-molecule kinetics and super-resolution microscopy by fluorescence imaging of transient binding on DNA origami." In: *Nano Lett.* 10.11, pp. 4756–4761.

Kay, A. J. and C. P. Hunter (2001). "CDC-42 regulates PAR protein localization and function to control cellular and embryonic polarity in C. elegans". In: *Curr. Biol.*

Kemphues, K. J., J. R. Priess, D. G. Morton, and N. Cheng (1988). "Identification of genes required for cytoplasmic localization in early C. elegans embryos". In: *Cell.*

Keppler, A., S. Gendreizig, T. Gronemeyer, H. Pick, H. Vogel, and K. Johnsson (2003). "A general method for the covalent labeling of fusion proteins with small molecules in vivo". In: *Nat. Biotechnol.*

Khalili-Araghi, F., J. Gumbart, P. C. Wen, M. Sotomayor, E. Tajkhorshid, and K. Schulten (2009). *Molecular dynamics simulations of membrane channels and transporters.*

Kiuchi, T., M. Higuchi, A. Takamura, M. Maruoka, and N. Watanabe (July 2015). "Multitarget super-resolution microscopy with high-density labeling by exchangeable probes". In: *Nat. Methods* 12.8, pp. 743–746.

Kono, K., S. Yoshiura, I. Fujita, Y. Okada, A. Shitamukai, T. Shibata, and F. Matsuzaki (2019). "Reconstruction of Par-dependent polarity in apolar cells reveals a dynamic process of cortical polarization". In: *Elife.*

Krahn, M. P., J. Bückers, L. Kastrup, and A. Wodarz (Sept. 2010). "Formation of a Bazooka-Stardust complex is essential for plasma membrane polarity in epithelia." In: *J. Cell Biol.* 190.5, pp. 751–760.

Kumfer, K. T., S. J. Cook, J. M. Squirrell, K. W. Eliceiri, N. Peel, K. F. O'Connell, and J. G. White (2010). "CGEF-1 and CHIN-1 regulate CDC-42 activity during asymmetric division in the Caenorhabditis elegans embryo". In: *Mol. Biol. Cell.*

Lang, C. F. and E. Munro (Oct. 2017). "The PAR proteins: from molecular circuits to dynamic self-stabilizing cell polarity." In: *Development* 144.19, pp. 3405–3416.

Laprise, P., S. Beronja, N. F. Silva-Gagliardi, M. Pellikka, A. M. Jensen, C. J. McGlade, and U. Tepass (2006). "The FERM Protein Yurt Is a Negative Regulatory Component of the Crumbs Complex that Controls Epithelial Polarity and Apical Membrane Size". In: *Dev. Cell*.

Lee, C. Y., K. J. Robinson, and C. Q. Doe (2006). "Lgl, Pins and aPKC regulate neuroblast self-renewal versus differentiation". In: *Nature*.

Lee, J. L. and C. H. Streuli (2014). *Integrins and epithelial cell polarity*.

Lemmon, M. A. (2008). *Membrane recognition by phospholipid-binding domains*.

Letizia, A., S. Ricardo, B. Moussian, N. Martín, and M. Llimargas (May 2013). "A functional role of the extracellular domain of Crumbs in cell architecture and apicobasal polarity." In: *J. Cell Sci.* 126.Pt 10, pp. 2157–2163.

Li, R. and B. Bowerman (Mar. 2010). "Symmetry breaking in biology." In: *Cold Spring Harb. Perspect. Biol.* 2.3, a003475–a003475.

Lin, D., A. S. Edwards, J. P. Fawcett, G. Mbamalu, J. D. Scott, and T. Pawson (Aug. 2000). "A mammalian PAR-3-PAR-6 complex implicated in Cdc42/Rac1 and aPKC signalling and cell polarity." In: *Nat. Cell Biol.* 2.8, pp. 540–547.

Lin, Y.-H., H. Currinn, S. M. Pocha, A. Rothnie, T. Wassmer, and E. Knust (Dec. 2015). "AP-2-complex-mediated endocytosis of Drosophila Crumbs regulates polarity by antagonizing Stardust." In: *J. Cell Sci.* 128.24, pp. 4538–4549.

Lin, M. Y., H. M. Lindsay, D. A. Weitz, R. C. Ball, R. Klein, and P. Meakin (1989). "Universality in colloid aggregation". In: *Nature*.

Lin, R., A. H. Clowsley, T. Lutz, D. Baddeley, and C. Soeller (May 2019). "3D super-resolution microscopy performance and quantitative analysis assessment using DNA-PAINT and DNA origami test samples." In: *Methods*.

Lindahl, E. and M. S. Sansom (2008). *Membrane proteins: molecular dynamics simulations*.

Los, G. V. et al. (2008). "HaloTag: A novel protein labeling technology for cell imaging and protein analysis". In: *ACS Chem. Biol.*

Lovegrove, H. E., D. T. Bergstralh, and D. St Johnston (2019). "The role of integrins in Drosophila egg chamber morphogenesis". In: *bioRxiv*.

Lutzmann, M., R. Kunze, A. Buerer, U. Aebi, and E. Hurt (2002). "Modular self-assembly of a Y-shaped multiprotein complex from seven nucleoporins". In: *EMBO J.*

Mah, I. K., R. Soloff, S. M. Hedrick, and F. V. Mariani (2015). “Atypical PKC-iota controls stem cell expansion via regulation of the Notch pathway”. In: *Stem Cell Reports*.

Makarova, O., M. H. Roh, C.-J. Liu, S. Laurinec, and B. Margolis (Jan. 2003). “Mammalian Crumbs3 is a small transmembrane protein linked to protein associated with Lin-7 (Pals1).” In: *Gene* 302.1-2, pp. 21–29.

Mao, Y. S., B. Zhang, and D. L. Spector (2011). *Biogenesis and function of nuclear bodies*.

Markova, O., J. Alberts, E. Munro, and P. F. Lenne (2012). “Bond flexibility and low valence promote finite clusters of self-aggregating particles”. In: *Phys. Rev. Lett.*

Martins Ratamero, E. and S. J. Royle (2019). “Calculating the maximum capacity of intracellular transport vesicles”. In: *bioRxiv*.

McCaffrey, L. M. and I. G. Macara (2009). “The Par3/aPKC interaction is essential for end bud remodeling and progenitor differentiation during mammary gland morphogenesis”. In: *Genes Dev.*

Mcgorty, R., D. Kamiyama, and B. Huang (2013). “Active microscope stabilization in three dimensions using image correlation”. In: *Opt. Nanoscopy*.

McKinley, R. F. A. and T. J. C. Harris (2012). “Displacement of basolateral Bazooka/PAR-3 by regulated transport and dispersion during epithelial polarization in Drosophila”. In: *Mol. Biol. Cell*.

Meakin, P., T. Vicsek, and F. Family (1985). “Dynamic cluster-size distribution in cluster-cluster aggregation: Effects of cluster diffusivity”. In: *Phys. Rev. B*.

Médina, E., J. Williams, E. Klipfell, D. Zarnescu, G. Thomas, and A. Le Bivic (Sept. 2002). “Crumbs interacts with moesin and beta(Heavy)-spectrin in the apical membrane skeleton of Drosophila.” In: *J. Cell Biol.* 158.5, pp. 941–951.

Mizuno, K., A. Suzuki, T. Hirose, K. Kitamura, K. Kutsuzawa, M. Futaki, Y. Amano, and S. Ohno (2003). “Self-association of PAR-3-mediated by the conserved N-terminal domain contributes to the development of epithelial tight junctions”. In: *J. Biol. Chem.*

Morais-de-Sá, E., V. Mirouse, and D. St Johnston (Apr. 2010). “aPKC Phosphorylation of Bazooka Defines the Apical/Lateral Border in Drosophila Epithelial Cells”. In: *Cell* 141.3, pp. 509–523.

Morais-De-Sá, E. and C. Sunkel (2013). “Adherens junctions determine the apical position of the midbody during follicular epithelial cell division”. In: *EMBO Rep.*

Motegi, F., S. Zonies, Y. Hao, A. A. Cuenca, E. Griffin, and G. Seydoux (2011). “Microtubules induce self-organization of polarized PAR domains in Caenorhabditis elegans zygotes”. In: *Nat. Cell Biol.*

Munro, E., J. Nance, and J. R. Priess (2004). “Cortical flows powered by asymmetrical contraction transport PAR proteins to establish and maintain anterior-posterior polarity in the early C. elegans embryo”. In: *Dev. Cell*.

Nakayama, M., T. M. Goto, M. Sugimoto, T. Nishimura, T. Shinagawa, S. Ohno, M. Amano, and K. Kaibuchi (Feb. 2008). “Rho-kinase phosphorylates PAR-3 and disrupts PAR complex formation.” In: *Dev. Cell* 14.2, pp. 205–215.

Nam, S. C. and K. W. Choi (2003). “Interaction of Par-6 and Crumbs complexes is essential for photoreceptor morphogenesis in Drosophila”. In: *Development*.

Nashchekin, D., A. R. Fernandes, and D. St Johnston (July 2016). “Patronin/Shot Cortical Foci Assemble the Noncentrosomal Microtubule Array that Specifies the Drosophila Anterior-Posterior Axis”. In: *Dev. Cell* 38.1, pp. 61–72.

Nishimura, A. and M. E. Linder (2013). “Identification of a Novel Prenyl and Palmitoyl Modification at the CaaX Motif of Cdc42 That Regulates RhoGDI Binding”. In: *Mol. Cell. Biol.*

Nishio, M. et al. (2007). “Control of cell polarity and motility by the PtdIns(3,4,5)P3 phosphatase SHIP1”. In: *Nat. Cell Biol.*

Ori, A. et al. (2013). “Cell type-specific nuclear pores: A case in point for context-dependent stoichiometry of molecular machines”. In: *Mol. Syst. Biol.*

Park, S., W. Kang, Y.-D. Kwon, J. Shim, S. Kim, B.-K. Kaang, and S. Hohng (Mar. 2018). “Superresolution fluorescence microscopy for 3D reconstruction of thick samples.” In: *Mol. Brain* 11.1, p. 17.

Pawley, J. B. (2006). *Handbook of biological confocal microscopy: Third edition*.

Peruani, F., A. Deutsch, and M. Bär (2006). “Nonequilibrium clustering of self-propelled rods”. In: *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*

Petronczki, M. and J. A. Knoblich (2001). “DmPAR-6 directs epithelial polarity and asymetric cell division of neuroblasts in Drosophila”. In: *Nat. Cell Biol.*

Pollard, T. D. (2010). “A Guide to Simple and Informative Binding Assays”. In: *Mol. Biol. Cell*.

Port, F., H. M. Chen, T. Lee, and S. L. Bullock (2014). “Optimized CRISPR/Cas tools for efficient germline and somatic genome engineering in Drosophila”. In: *Proc. Natl. Acad. Sci. U. S. A.*

Recouvreux, P. and P.-F. Lenne (Feb. 2016). “Molecular clustering in the cell: from weak interactions to optimized functional architectures.” In: *Curr. Opin. Cell Biol.* 38, pp. 18–23.

Remorino, A., S. De Beco, F. Cayrac, F. Di Federico, G. Cornilleau, A. Gautreau, M. C. Parrini, J. B. Masson, M. Dahan, and M. Coppey (2017). "Gradients of Rac1 Nanoclusters Support Spatial Patterns of Rac1 Signaling". In: *Cell Rep.*

Rodriguez, J. et al. (Aug. 2017). "aPKC Cycles between Functionally Distinct PAR Protein Assemblies to Drive Cell Polarity." In: *Dev. Cell* 42.4, 400–415.e9.

Röper, K. (Nov. 2012). "Anisotropy of Crumbs and aPKC drives myosin cable assembly during tube formation." In: *Dev. Cell* 23.5, pp. 939–953.

Rubin-Delanchy, P., G. L. Burn, J. Griffié, D. J. Williamson, N. A. Heard, A. P. Cope, and D. M. Owen (2015). "Bayesian cluster identification in single-molecule localization microscopy data". In: *Nat. Methods*.

Rust, M. J., M. Bates, and X. Zhuang (Oct. 2006). "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM)." In: *Nat. Methods* 3.10, pp. 793–795.

Sailer, A., A. Anneken, Y. Li, S. Lee, and E. Munro (Oct. 2015). "Dynamic Opposition of Clustered Proteins Stabilizes Cortical Polarity in the C. elegans Zygote." In: *Dev. Cell* 35.1, pp. 131–142.

Schlichthaerle, T. et al. (2019). "Direct Visualization of Single Nuclear Pore Complex Proteins Using Genetically-Encoded Probes for DNA-PAINT". In: *Angew. Chemie Int. Ed.*

Schmidt, A., Z. Lv, and J. Grosshans (Jan. 2018). "ELMO and Sponge specify subapical restriction of Canoe and formation of the subapical domain in early Drosophila embryos". In: *Development* 145.2, dev157909.

Schneider, M., A. A. Khalil, J. Poulton, C. Castillejo-Lopez, D. Egger-Adam, A. Wodarz, W.-M. Deng, and S. Baumgartner (Oct. 2006). "Perlecan and Dystroglycan act at the basal side of the Drosophila follicular epithelium to maintain epithelial organization." In: *Development* 133.19, pp. 3805–3815.

Schnitzbauer, J., M. T. Strauss, T. Schlichthaerle, F. Schueder, and R. Jungmann (June 2017). "Super-resolution microscopy with DNA-PAINT." In: *Nat. Protoc.* 12.6, pp. 1198–1228.

Schueder, F., J. Lara-Gutiérrez, B. J. Beliveau, S. K. Saka, H. M. Sasaki, J. B. Woehrstein, M. T. Strauss, H. Grabmayr, P. Yin, and R. Jungmann (2017). "Multiplexed 3D super-resolution imaging of whole cells using spinning disk confocal microscopy and DNA-PAINT". In: *Nat. Commun.*

Semplice, M., A. Veglio, G. Naldi, G. Serini, and A. Gamba (2012). "A bistable model of cell polarity". In: *PLoS One*.

Sengupta, P., T. Jovanovic-Talisman, D. Skoko, M. Renz, S. L. Veatch, and J. Lippincott-Schwartz (2011). "Probing protein heterogeneity in the plasma membrane using PALM and pair correlation analysis". In: *Nat. Methods*.

Seydoux, G. (2018). *The P Granules of C. elegans: A Genetic Model for the Study of RNA–Protein Condensates*.

Shahab, J., M. D. Tiwari, M. Honemann-Capito, M. P. Krahn, and A. Wodarz (2015). "Bazooka/PAR3 is dispensable for polarity in drosophila follicular epithelial cells". In: *Biol. Open*.

Shan, Z., Y. Tu, Y. Yang, Z. Liu, M. Zeng, H. Xu, J. Long, M. Zhang, Y. Cai, and W. Wen (Feb. 2018). "Basal condensation of Numb and Pon complex via phase transition during Drosophila neuroblast asymmetric division." In: *Nat. Commun.* 9.1, p. 737.

Sharma, P., R. Varma, R. C. Sarasij, Ira, K. Gousset, G. Krishnamoorthy, M. Rao, and S. Mayor (2004). "Nanoscale Organization of Multiple GPI-Anchored Proteins in Living Cell Membranes". In: *Cell*.

Sherman, E. et al. (2011). "Functional nanoscale organization of signaling molecules downstream of the T cell antigen receptor". In: *Immunity*.

Sherrard, K. M. and R. G. Fehon (May 2015). "The transmembrane protein Crumbs displays complex dynamics during follicular morphogenesis and is regulated competitively by Moesin and aPKC". In: *Development* 142.10, pp. 1869–1878.

Sieber, J. J. et al. (Aug. 2007). "Anatomy and Dynamics of a Supramolecular Membrane Protein Cluster". In: *Science* 317.5841, pp. 1072–1076.

Simons, K. and S. D. Fuller (1985). *Cell surface polarity in epithelia.*

Soriano, E. V. et al. (2016). "aPKC Inhibition by Par3 CR3 Flanking Regions Controls Substrate Access and Underpins Apical-Junctional Polarization". In: *Dev. Cell*.

Sotillos, S., M. T. Díaz-Meco, E. Caminero, J. Moscat, and S. Campuzano (Aug. 2004). "DaPKC-dependent phosphorylation of Crumbs is required for epithelial cell polarity in Drosophila." In: *J. Cell Biol.* 166.4, pp. 549–557.

Spradling, A. C. (1993). "Developmental genetics of oogenesis". In: *Dev. Drosoph. melanogaster*.

Stein, J., F. Stehr, P. Schueler, P. Blumhardt, F. Schueder, J. Mücksch, R. Jungmann, and P. Schwille (2019). "Towards absolute molecular numbers in DNA-PAINT". In: *Nano Lett.*

Stein, W. von, A. Ramrath, A. Grimm, M. Müller-Borg, and A. Wodarz (2005). "Direct association of Bazooka/PAR-3 with the lipid phosphatase PTEN reveals a link between the PAR/aPKC complex and phosphoinositide signaling". In: *Development*.

Stone, M. B., S. A. Shelby, M. F. Núñez, K. Wisser, and S. L. Veatch (Feb. 2017). "Protein sorting by lipid phase-like domains supports emergent signaling function in B lymphocyte plasma membranes." In: *Elife* 6, p. 212.

Strutt, H., P. F. Langton, N. Pearson, K. J. McMillan, D. Strutt, and P. J. Cullen (2019). "Retromer Controls Planar Polarity Protein Levels and Asymmetric Localization at Intercellular Junctions". In: *Curr. Biol.*

Strutt, H., S. J. Warrington, and D. Strutt (2011). "Dynamics of Core Planar Polarity Protein Turnover and Stable Assembly into Discrete Membrane Subdomains". In: *Dev. Cell*.

Su, X., J. A. Ditlev, E. Hui, W. Xing, S. Banjade, J. Okrut, D. S. King, J. Taunton, M. K. Rosen, and R. D. Vale (2016). "Phase separation of signaling molecules promotes T cell receptor signal transduction". In: *Science (80-.).*

Suzuki, A. and S. Ohno (2006). "The PAR-aPKC system: Lessons in polarity". In: *J. Cell Sci.*

Tabuse, Y., Y. Izumi, F. Piano, K. J. Kemphues, J. Miwa, and S. Ohno (1998). "Atypical protein kinase C cooperates with PAR-3 to establish embryonic polarity in Caenorhabditis elegans". In: *Development*.

Tanentzapf, G., C. Smith, J. McGlade, and U. Tepass (Nov. 2000). "Apical, lateral, and basal polarization cues contribute to the development of the follicular epithelium during Drosophila oogenesis." In: *J. Cell Biol.* 151.4, pp. 891–904.

Tanentzapf, G. and U. Tepass (2003). "Interactions between the crumbs, lethal giant larvae and bazooka pathways in epithelial polarization". In: *Nat. Cell Biol.*

Tang, A. H., H. Chen, T. P. Li, S. R. Metzbower, H. D. MacGillavry, and T. A. Blanpied (2016). "A trans-synaptic nanocolumn aligns neurotransmitter release to receptors". In: *Nature*.

Tepass, U., C. Theres, and E. Knust (June 1990). "crumbs encodes an EGF-like protein expressed on apical membranes of Drosophila epithelial cells and required for organization of epithelia." In: *Cell* 61.5, pp. 787–799.

Tepass, U. (1996). "Crumbs, a component of the apical membrane, is required for zonula adherens formation in primary epithelia of Drosophila". In: *Dev. Biol.*

— (Jan. 2012). “The apical polarity protein network in Drosophila epithelial cells: regulation of polarity, junctions, morphogenesis, cell growth, and survival.” In: *Annu. Rev. Cell Dev. Biol.* 28.1, pp. 655–685.

Thevathasan, J. V. et al. (2019). “Nuclear pores as versatile reference standards for quantitative superresolution microscopy”. In: *Nat. Methods*.

Thompson, B. J. and N. Q. McDonald (2019). *Competitive Inhibition of aPKC by Par-3/Bazooka and Other Substrates*.

Thompson, R. E., D. R. Larson, and W. W. Webb (May 2002). “Precise nanometer localization analysis for individual fluorescent probes.” In: *Biophys. J.* 82.5, pp. 2775–2783.

Truong Quang, B.-A. and P.-F. Lenne (Jan. 2014). “Membrane microdomains: from seeing to understanding.” In: *Front. Plant Sci.* 5, p. 18.

Truong Quang, B.-A., M. Mani, O. Markova, T. Lecuit, and P.-F. Lenne (Nov. 2013). “Principles of E-cadherin supramolecular organization in vivo.” In: *Curr. Biol.* 23.22, pp. 2197–2207.

Tyagi, S. and F. R. Kramer (1996). “Molecular Beacons: Probes that Fluoresce Upon Hybridization”. In: *Nat. Biotechnol.*

Vangindertael, J., R. Camacho, W. Sempels, H. Mizuno, P. Dedecker, and K. P. F. Janssen (Mar. 2018). “An introduction to optical super-resolution microscopy for the adventurous biologist.” In: *Methods Appl. Fluoresc.* 6.2, p. 22003.

Vogelsang, J., C. Steinhauer, C. Forthmann, I. H. Stein, B. Person-Skegro, T. Cordes, and P. Tinnefeld (2010). “Make them blink: Probes for super-resolution microscopy”. In: *ChemPhysChem*.

Wade, O. K. et al. (2019). “124-Color Super-resolution Imaging by Engineering DNA-PAINT Blinking Kinetics”. In: *Nano Lett.*

Walther, R. F. and F. Pichaud (June 2010). “Crumbs/DaPKC-dependent apical exclusion of Bazooka promotes photoreceptor polarity remodeling.” In: *Curr. Biol.* 20.12, pp. 1065–1074.

Wang, J. T., J. Smith, B.-C. Chen, H. Schmidt, D. Rasoloson, A. Paix, B. G. Lambrus, D. Calidas, E. Betzig, and G. Seydoux (Dec. 2014). “Regulation of RNA granule dynamics by phosphorylation of serine-rich, intrinsically disordered proteins in C. elegans.” In: *Elife* 3, e04591.

Wang, Q., T. W. Hurd, and B. Margolis (2004). "Tight junction protein Par6 interacts with an evolutionarily conserved region in the amino terminus of PALS1/stardust". In: *J. Biol. Chem.*

Wang, S.-C., T. Y. F. Low, Y. Nishimura, L. Gole, W. Yu, and F. Motegi (July 2017). "Cortical forces and CDC-42 control clustering of PAR proteins for Caenorhabditis elegans embryonic polarization". In: *Nat. Cell Biol.* 126, p. 127.

Watts, J. L., B. Etemad-Moghadam, S. Guo, L. Boyd, B. W. Draper, C. C. Mello, J. R. Priess, and K. J. Kemphues (Oct. 1996). "par-6, a gene involved in the establishment of asymmetry in early C. elegans embryos, mediates the asymmetric localization of PAR-3." In: *Development* 122.10, pp. 3133–3140.

Weber, C., T. Michaels, and L. Mahadevan (2019). "Spatial control of irreversible protein aggregation". In: *Elife*.

Weberruss, M. and W. Antonin (Dec. 2016). "Perforating the nuclear boundary - how nuclear pore complexes assemble." In: *J. Cell Sci.* 129.24, pp. 4439–4447.

Whelan, D. R. and T. D. M. Bell (Jan. 2015). "Image artifacts in single molecule localization microscopy: why optimization of sample preparation protocols matters." In: *Sci. Rep.* 5, p. 7924.

Whitney, D. S., F. C. Peterson, A. W. Kittell, J. M. Egner, K. E. Prehoda, and B. F. Volkman (Mar. 2016). "Binding of Crumbs to the Par-6 CRIB-PDZ Module Is Regulated by Cdc42." In: *Biochemistry* 55.10, pp. 1455–1461.

Wicksell, S. D. (1925). "The Corpuscle Problem: A Mathematical Study of a Biometric Problem". In: *Biometrika*.

Williamson, D. J., G. L. Burn, J. Griffié, D. M. Davis, and D. M. Owen (Dec. 2018). "Machine-learning for cluster analysis of localization microscopy data." In: *bioRxiv*, p. 505719.

Wirtz-Peitz, F., T. Nishimura, and J. A. Knoblich (2008). "Linking Cell Cycle to Asymmetric Division: Aurora-A Phosphorylates the Par Complex to Regulate Numb Localization". In: *Cell*.

Wit, E., E. van den Heuvel, and J. W. Romeijn (2012). "'All models are wrong. ': An introduction to model uncertainty". In: *Stat. Neerl.*

Wodarz, A., A. Ramrath, A. Grimm, and E. Knust (Sept. 2000). "Drosophila atypical protein kinase C associates with Bazooka and controls polarity of epithelia and neuroblasts." In: *J. Cell Biol.* 150.6, pp. 1361–1374.

Woodhams, S. G., R. Markus, P. R. Gowler, T. J. Self, and V. Chapman (2019). "Cell type–specific super-resolution imaging reveals an increase in calcium-permeable AMPA receptors at spinal peptidergic terminals as an anatomical correlate of inflammatory pain". In: *Pain*.

Wu, M. (Dec. 2015). "Shaping Developing Tissues with Light." In: *Dev. Cell* 35.5, pp. 533–534.

Xiao, Z., J. Patrakka, M. Nukui, L. Chi, D. Niu, C. Betsholtz, T. Pikkarainan, S. Vainio, and K. Tryggvason (2011). "Deficiency in crumbs homolog 2 (Crb2) affects gastrulation and results in embryonic lethality in mice". In: *Dev. Dyn.*

Xu, T. and G. M. Rubin (Apr. 1993). "Analysis of genetic mosaics in developing and adult Drosophila tissues." In: *Development* 117.4, pp. 1223–1237.

Xue, B., K. Krishnamurthy, D. C. Allred, and S. K. Muthuswamy (2013). "Loss of Par3 promotes breast cancer metastasis by compromising cell-cell cohesion". In: *Nat. Cell Biol.*

Yamanaka, T. et al. (2001). "PAR-6 regulates aPKC activity in a novel way and mediates cell-cell contact-induces formation of the epithelial junctional complex". In: *Genes to Cells*.

Yamanaka, T., Y. Horikoshi, N. Izumi, A. Suzuki, K. Mizuno, and S. Ohno (2006). "Lgl mediates apical domain disassembly by suppressing the PAR-3-aPKC-PAR-6 complex to orient apical membrane polarity". In: *J. Cell Sci.*

Zihni, C., C. Mills, K. Matter, and M. S. Balda (2016). *Tight junctions: From simple barriers to multifunctional molecular gates*.

Zou, J., X. Wang, and X. Wei (June 2012). "Crb Apical Polarity Proteins Maintain Zebrafish Retinal Cone Mosaics via Intercellular Binding of Their Extracellular Domains". In: *Dev. Cell* 22.6, pp. 1261–1274.

Appendices

Appendix A

Primers used for cloning and CRISPR target sequences

Primer Name	**Sequence**
Crb2F	AGGGCACACAACAGCAAAC
Crb2R	GACTGCGACTACCATGTGCA
HaloCrumbsF	CGGCCAAGGAGGACGGGCCTGGAGGTAGTGCAGAAATCGGTACTGGCTTTC
HaloCrumbsR	ATGGCAATGTCTGTGGTCGAGCTTCCACCGCCGGAAATCTCGAGCGT
CrumbsTagR	AGGCCCGTCCTCCTTGGC
CrumbsTagF	TCGACCACAGACATTGCCATCA
Par6Kpn1p	ACGTGGTACCTGGTTCGGTTTCGGTTCGC
Par6Not1m	ACGTGCGGCCGCGCAAGCAAAAGAGTGGTCG
Par6CRF1mutF	GCAGCCTCGAACGCCTCCACGATAATGGCC
Par6CRF1mutR	TATCGTGGAGGCGTTCGAGGCTGCCTGCTG
pBS-Par6-vec-rev	CGATCCCAAATGCAGCAC
pBS-Par6-vec-fwd	TAGGATTAACCCGGCGATAC
Halo-Tag-fwd	GCTGCATTTGGGATCGGCAGAAATCGGTACTGGC
Halo-Tag-rev	CGCCGGGTTAATCCTAGCCGGAAATCTCGAGCGT
aPKC5UT1p	CTCCGCTTTGTGCTCTCCTTCC
aPKCint2m	AGTCGCAGGAAAACAGACGC

aPKCCRF1mut	CGACGGCAGCAGCGTCAGCTTGAATAGCGCtAGCATGAAtATGGCCAATACGCCCAAT
GFPaPKCfus2m	GCTGACGCTGCTGCCGTCGTTCAGAATTTGCGAGGGCATTTTCTGGGAACCGCTTCCCTTGTACAGCTCGTCCATG
GFPaPKCfus1p	TTACCCAGCTAGCAAATGGTGAGCAAGGGCGAGGAG
GFPaPKCfus1m	GCCCTTGCTCACCATTTGCTAGCTGGGTAAAAT
aPKCnt-rev	TTGCTAGCTGGGTAAAATATTTTGATATCACG
SNAPaPKC-fwd	ATTTTACCCAGCTAGCAAATGACTAGTGACAAAGACTGCGAAATG
SNAPaPKC-rev	CTGGGAACCGCTTCCACCCAGCCCAGGCTTGCC
aPKCnt-fwd	GGAAGCGGTTCCCAGAAAATGC
Halo-aPKC-fwd	TTTTACCCAGCTAGCAAATGGCAGAAATCGGTACTGG
Halo-aPKC-rev	CTGGGAACCGCTTCCGCCGGAAATCTCGAGCGT
Nup160Halo-F1	AACGTTGCTCCTCGCTGTTA
Nup160Halo-R1	ACGTTCTTCTTGCAGATCGGTTCGCATCTG
Nup160Halo-F2	TAATAGTTCGTTTAAATCAGTTTATTTAATAAG
Nup160Halo-R2	GATCAAGCGGAGAACCTACG
Nup188SNAP-F	TCTCCGCCTGCAATGAGTAA
Nup188SNAP-R	GTCGCAATGGGGCATTACATT

Table A.1: Primers used for cloning.

Gene	**Target sequence (PAM in red)**
crumbs	GGTCGAAGGCCCGTCCTCCTTTG
par6	GGCCATTATCGTGGACGCATTGG
aPKC	GAATAGCGCCAGTATGAACATGG
nup160 1	GGCCGTTCTGCAGATGCGAACGG
nup160 2	GCAGATGCGAACGGATCTTCAGG
nup188 1	TTCGATTGTAAAGTAATTCATGG

Table A.2: CRISPR target sequences.

Appendix B

Integrated signal anisotropy analysis

Author: Richard Butler

```
#       Line Profile Analysis
#       Calculates line intensity profile spectral density as the
    Fourier transform
#       of the autocorrelation to measure spectral power, principal
    frequency and
#       signal anisotropy.
#
#       Copyright (C) 2019 Richard Butler, Gurdon Institute Imaging
    Facility
#
#       This program is free software: you can redistribute it and/or
     modify
#       it under the terms of the GNU General Public License as
    published by
#       the Free Software Foundation, either version 3 of the License
    , or
#       (at your option) any later version.
#
#       This program is distributed in the hope that it will be
    useful,
#       but WITHOUT ANY WARRANTY; without even the implied warranty
    of
#       MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
```

```
#       GNU General Public License for more details.

#       You should have received a copy of the GNU General Public
    License
#       along with this program.  If not, see <https://www.gnu.org/
    licenses/>.


import sys, re, random, itertools
import math as maths

from ij import IJ, WindowManager, Prefs, ImagePlus, ImageStack
from ij.plugin import ImageCalculator, Duplicator, Straightener
from ij.plugin.filter import GaussianBlur, ThresholdToSelection,
    MaximumFinder
from ij.process import ImageProcessor, FloatProcessor,
    ImageStatistics, Blitter, FHT, AutoThresholder, FloatPolygon
from ij.measure import Measurements, ResultsTable
from ij.gui import Roi, ShapeRoi, Line, Overlay

from org.jfree.chart import JFreeChart, ChartFactory, ChartPanel,
    ChartFrame, LegendItemCollection, LegendItem
from org.jfree.chart.plot import PlotOrientation
from org.jfree.chart.axis import LogAxis
from org.jfree.chart.annotations import XYLineAnnotation
from org.jfree.chart.renderer.xy import StandardXYBarPainter
from org.jfree.chart.annotations import XYAnnotation
from org.jfree.data.xy import DefaultXYDataset, XYSeries,
    XYSeriesCollection
from org.jfree.chart.renderer.xy import XYLineAndShapeRenderer

from java.awt import Color, BasicStroke, BorderLayout
from javax.swing import JFrame


def PCC(valuesA, valuesB):
```

```
        sumA = sum( valuesA )
        sumB = sum( valuesB )
        covar = 0
        varA = 0
        varB = 0
        n = min(len(valuesA),len(valuesB))
        meanA = sumA/n
        meanB = sumB/n
        for i in range(n):
                covar += (valuesA[i]-meanA)*(valuesB[i]-meanB)
                varA += (valuesA[i]-meanA)*(valuesA[i]-meanA)
                varB += (valuesB[i]-meanB)*(valuesB[i]-meanB)
        pcc = covar/maths.sqrt(varA*varB)
        return pcc

def plotLine(dataset, title, x, y, annotations):
        chart = ChartFactory.createScatterPlot( title, x, y, dataset,
            PlotOrientation.VERTICAL, False,True,False )
        plot = chart.getPlot()

        #yAxis = LogAxis(y)
        #yAxis.setBase(maths.e)
        #plot.setRangeAxis( yAxis )

        plot.setBackgroundPaint(Color.PINK)
        plot.setDomainGridlinePaint(Color.GRAY)
        plot.setRangeGridlinePaint(Color.GRAY)
        if annotations is not None:
                for ann in annotations:
                        plot.addAnnotation(ann)

        renderer = plot.getRenderer()
        n = dataset.getSeriesCount()
        for s in range(n):
                renderer.setSeriesLinesVisible(s, True)
                renderer.setSeriesShapesVisible(s, False)
```

```
                renderer.setSeriesStroke(s, BasicStroke(1.0))
                renderer.setSeriesPaint(s, Color.BLUE)

        chartPanel = ChartPanel(chart)
        frame = JFrame(title)
        frame.setLayout(BorderLayout())
        frame.setSize(800, 800)
        frame.setLocationRelativeTo(None)
        frame.add(chartPanel, BorderLayout.CENTER)
        frame.pack()
        frame.setVisible(True)

def autocorrelation(imp, width):
        roi = imp.getRoi()
        if roi is None or not roi.isLine():
                IJ.error("Line selection required.")
                return
        lineip = Straightener().straighten(imp, roi, width)
        lineip.blurGaussian(2)
        W = lineip.getWidth()
        W2 = W/float(2)
        H = lineip.getHeight()
        profile = [[],[]]
        for x in range(W):
                val = 0
                for y in range(H):
                        val += lineip.getf(x,y)
                val /= H
                profile[0].append(x*cal.pixelWidth)
                profile[1].append(val)

        profileMean = sum(profile[1])/len(profile[1])
        ssd = sum([ pow(p-profileMean,2) for p in profile[1] ])
        profileStdDev = maths.sqrt( ssd/len(profile[1]) )

        autoCorrelation = [[],[]]
```

```
#for r in range(-(W/2),(W/2)+1):
for r in range(int(W2)):
        offset = profile[1][r:] + profile[1][:r] #rotated
        pcc = PCC(profile[1], offset)
        autoCorrelation[0].append(r*cal.pixelWidth)
        autoCorrelation[1].append(pcc)
profileDataset = DefaultXYDataset()
acDataset = DefaultXYDataset()
profileDataset.addSeries("profile", profile)
acDataset.addSeries("ac", autoCorrelation)


maximi = MaximumFinder.findMaxima(autoCorrelation[1], 0.1,
    False)
maximx = [ autoCorrelation[0][index] for index in maximi ]
maximy = [ autoCorrelation[1][index] for index in maximi ]
peakAnn = []
nPeaks = 0
for m in range(1,len(maximi)):  #don't include maximum
    amplitude (R=1) peak at offset 0
        if maximy[m] > 0.0:
                peakAnn.append( XYLineAnnotation(maximx[m],
                    -10, maximx[m], 10, BasicStroke(2), Color.
                    RED) )
                nPeaks += 1
lamb = 0.0
if nPeaks > 0:
        lamb = (W2*cal.pixelWidth)/float(nPeaks) #wavelength,
            distance units
nu = nPeaks/(W2*cal.pixelWidth)                 #wavenumber,
    1/lamb, reciprocal units

#FFT of Autocorrelation = Spectral Density
acfft = [[],[]]
acfft[1] = [f for f in FHT().fourier1D(autoCorrelation[1],
    FHT.HAMMING) ]
```

```
        acfft[0] = [p*cal.pixelWidth for p in range(len(acfft[1]))]
        dataset = DefaultXYDataset()
        dataset.addSeries("acfft", acfft)
        power = sum(acfft[1])    #sum of spectral densities
        maxf = acfft[0][ acfft[1].index(max(acfft[1])) ]          #
            frequency at maximum amplitude

        rt = ResultsTable.getResultsTable()
        row = rt.getCounter()
        rt.setValue("Image", row, imp.getTitle())
        rt.setValue("Length␣("+u"\u00b5"+"m)", row, W*cal.pixelWidth)
        rt.setValue("Mean", row, profileMean)
        rt.setValue("StdDev", row, profileStdDev)
        rt.setValue("Cycles", row, nPeaks)
        rt.setValue(u"\u03bb"+"␣("+u"\u00b5"+"m)", row, lamb)
        rt.setValue(u"\u03bd"+"␣("+u"\u00b5"+"m"+"^-1)", row, nu)
        rt.setValue("Spectral␣Power", row, power)
        rt.setValue("Principal␣Frequency␣(c/"+u"\u00b5"+"m"+")", row,
             maxf)
        rt.setValue("Integrated␣Signal␣Anisotropy", row,
            profileStdDev*maxf)     #StdDev * Principal Frequency
        rt.show("Results")

        #plotLine(profileDataset, imp.getTitle()+" Line Profile", "
            Position ("+cal.getUnit()+")", "Intensity", None)
        #plotLine(dataset, imp.getTitle()+" Spectral Density : Power
            = "+IJ.d2s(power,2), "Frequency (c/"+cal.getUnit()+")", "
            Amplitude", None)
        #plotLine(acDataset, imp.getTitle()+" Autocorrelation : "+u"\
            u03bd"+" = "+IJ.d2s(nu,2)+" ("+u"\u00b5"+"m"+"^-1)", "
            Offset ("+cal.getUnit()+")", "R", peakAnn)
        #plotLine(dataset, imp.getTitle()+" FFT", "Frequency (cycles
            /"+cal.getUnit()+")", "Amplitude", None)

def fourier(imp, width):
        roi = imp.getRoi()
```

```
        if roi is None or not roi.isLine():
                IJ.error("Line selection required.")
                return
        lineip = Straightener().straighten(imp, roi, width)
        lineip.blurGaussian(1)
        W = lineip.getWidth()
        W2 = W/float(2)
        H = lineip.getHeight()
        profile = []
        for x in range(W):
                val = 0
                for y in range(H):
                        val += lineip.getf(x,y)
                val /= H
                profile.append(val)
        fft = [[],[]]

        fft[1] = [f for f in FHT().fourier1D(profile, FHT.HAMMING) ]
        fft[0] = [p*cal.pixelWidth for p in range(len(fft[1]))]
        dataset = DefaultXYDataset()
        dataset.addSeries("fft", fft)


imp = WindowManager.getCurrentImage()
W = imp.getWidth()
H = imp.getHeight()
imp.setOverlay(None)
global cal
cal = imp.getCalibration()
ol = Overlay()

autocorrelation(imp, 10)
#fourier(imp, 10)
```

Appendix C

Effective labelling efficiency analysis

Author: Richard Butler

```
#       Eightfold_Path
#       Measures pore labelling efficiency by segmenting pores
#       and dividing them into eight segments aligned to the maximum
    labelling
#       frequency. Efficiency is calculated as number of labelled
    segments
#       divided by eight.
#
#       Copyright (C) 2019 Richard Butler, Gurdon Institute Imaging
    Facility
#
#       This program is free software: you can redistribute it and/or
     modify
#       it under the terms of the GNU General Public License as
    published by
#       the Free Software Foundation, either version 3 of the License
    , or
#       (at your option) any later version.
#
#       This program is distributed in the hope that it will be
    useful,
#       but WITHOUT ANY WARRANTY; without even the implied warranty
    of
```

```
#       MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
#       GNU General Public License for more details.

#       You should have received a copy of the GNU General Public
    License
#       along with this program.  If not, see <https://www.gnu.org/
    licenses/>.


import csv
import math as maths

from ij import IJ, ImagePlus
from ij.plugin import Duplicator
from ij.plugin.filter import MaximumFinder, ThresholdToSelection, EDM
from ij.process import ImageProcessor, FloatProcessor, ByteProcessor,
     Blitter, AutoThresholder, FloatPolygon
from ij.measure import Calibration, ResultsTable
from ij.gui import Roi, ShapeRoi, Line, PointRoi, TextRoi, PolygonRoi
    , Overlay, WaitForUserDialog

from java.awt import Color, Dimension, BasicStroke, Font
from java.awt.geom import Ellipse2D

from org.jfree.chart import JFreeChart, ChartFactory, ChartPanel,
    ChartFrame
from org.jfree.chart.plot import PlotOrientation
from org.jfree.chart.annotations import XYLineAnnotation
from org.jfree.data.xy import DefaultXYDataset
from org.jfree.chart.axis import NumberTickUnit
from org.jfree.data.statistics import HistogramDataset, HistogramType
from org.jfree.chart.renderer.xy import StandardXYBarPainter


calibratedPixelSize = 160.0     #calibration to apply to coordinates
    (not image), nm
```

```
minPpc = 50       #min number of points per complex
minR = 60.0       #pore radius, nm
maxR = 140.0
minA = maths.pi*minR*minR #nm^2
maxA = maths.pi*maxR*maxR #nm^2
innerR = 30.0     #exclude detections within this radius of the pore
    centroid
minCirc = 0.75    #min pore circularity
showImage = True

segA = (2*maths.pi/8.0)
segDivs = [x * segA - maths.pi - (segA/2.0) for x in range(1,9)]

twopi = 2*maths.pi                #whole circle
qpi = (maths.pi/4.0)      #1/8th circle


def normAngle(angle):
        while angle>2*maths.pi:
                angle = 2*maths.pi - angle
        while angle<0:
                angle = angle + 2*maths.pi
        return angle

def getSegCounts(detections, centroid):
        angles = []
        for p in detections:
                radius = maths.sqrt( ((centroid[0]-p[0])*(centroid
                    [0]-p[0])) + ((centroid[1]-p[1])*(centroid[1]-p
                    [1])) ) * cal.pixelWidth
                if radius<innerR: continue       #don't include points
                     in central area
                theta = maths.atan2( centroid[1]-p[1], centroid[0]-p
                    [0] ) + maths.pi     #angle 0 .. 2PI
                angles.append( theta )
```

```
        #find the angle offset giving max frequency in one segment
        maxf = -1
        maxOffset = 0
        for offset in [0.05*i for i in range(16)]:       #0 .. pi/4
                freq = [0 for i in range(8)]
                for a in angles:
                        theta = a+offset
                        bindex = int( (theta/twopi) * (7) )
                        freq[bindex] += 1
                mf = max(freq)
                if mf > maxf:
                        maxf = mf
                        maxOffset = offset

        half = qpi/2.0
        segAngles = [ 0 for i in range(8) ]
        for s in range(8):
                segAngles[s] = s*qpi + maxOffset - half #max
                   frequency offset plus additional offset by half a
                   bin width
        counts = [0 for i in range(8)]

        for s in range(len(segAngles)): #assign bins
                for a in angles:
                        if a >= segAngles[s] and a <= segAngles[s]+
                           qpi:
                                counts[s] += 1

        return [segAngles, counts]

def getPoreRois(fp):
        fpMap = fp.duplicate()
        fpSub = fp.duplicate()
        sigma0 = minR #nm
        sigmaPx0 = sigma0/pixelSize
```

```
sigma1 = maxR #nm
sigmaPx1 = sigma1/pixelSize
fpMap.blurGaussian( sigmaPx0 )
fpSub.blurGaussian( sigmaPx1 )
fpMap.copyBits(fpSub, 0,0, Blitter.SUBTRACT)

at = AutoThresholder()
hist = fpMap.getHistogram(256)
stats = fpMap.getStatistics()
threshi = at.getThreshold( AutoThresholder.Method.Huang, hist
    )
thresh = stats.min + (threshi/float(255)) * (stats.max-stats.
   min)

mask = ByteProcessor(W, H)
for i in range(W*H):
        if fpMap.getf(i) >= thresh:
                mask.set(i, 255)
floatEdm = EDM().makeFloatEDM(mask, 0, False)
maxIp = MaximumFinder().findMaxima(floatEdm, 0.5,
   ImageProcessor.NO_THRESHOLD, MaximumFinder.SEGMENTED,
   False, True)
if (maxIp != None):
        mask.copyBits(maxIp, 0, 0, Blitter.AND)
mask.dilate()
mask.erode()

mask.setThreshold( 255, 255, ImageProcessor.NO_LUT_UPDATE )
composite = ThresholdToSelection().convert(mask)
rois = ShapeRoi(composite).getRois()
pores = []
for roi in rois:
        stats = roi.getStatistics()
        area = stats.area*cal.pixelWidth*cal.pixelHeight
        perim = roi.getLength()
        circ = 4*maths.pi*(stats.area/(perim*perim))
```

```
                if area>=minA and area<=maxA and circ>=minCirc:
                        pores.append(roi)
        return pores

path = IJ.getFilePath("CSV Coordinates...")
dataf = []
with open(path, 'r') as csvFile:
        minX = float('inf')
        minY = float('inf')
        maxX = float('-inf')
        maxY = float('-inf')
        for line in csv.reader(csvFile):

                #coordinates in camera pixel units
                x = float(line[0])
                y = float(line[1])

                #calibrate to nm
                x = x * calibratedPixelSize
                y = y * calibratedPixelSize

                minX = min(minX, x)
                minY = min(minY, y)
                maxX = max(maxX, x)
                maxY = max(maxY, y)
                dataf.append( [ x, y ] )

cal = Calibration()
cal.setUnit("nm")
pixelSize = 10.0
cal.pixelWidth = pixelSize
cal.pixelHeight = pixelSize
cal.pixelDepth  = 1.0

pad = 10
W = int( ((maxX-minX)/cal.pixelWidth)+(2*pad) )
```

```
H = int( ((maxY-minY)/cal.pixelHeight)+(2*pad) )

#make image for selection only
fp = FloatProcessor(W, H)
maxn = 0
for p in dataf:
        xi = int( pad + ((p[0]-minX)/cal.pixelWidth) )
        yi = int( pad + ((p[1]-minY)/cal.pixelHeight) )
        value = fp.getf(xi,yi)+1
        maxn = max(value, maxn)
        fp.setf( xi, yi, value )
imp = ImagePlus(path, fp)
imp.setDisplayRange(0, maxn/2)
imp.setCalibration(cal)
imp.show()

WaitForUserDialog("Select Area", "Select an area to analyse...\n(the
    whole area will be included if there is no selection)").show()
userRoi = None
if imp.getRoi() is not None:
        userRoi = imp.getRoi()

imp.close()

if userRoi is not None:
        minXnew = float('inf')
        minYnew = float('inf')
        maxXnew = float('-inf')
        maxYnew = float('-inf')
        keepf = []
        for d in dataf:
                xi = int( pad + ((d[0]-minX)/cal.pixelWidth) )
                yi = int( pad + ((d[1]-minY)/cal.pixelHeight) )
                if userRoi.contains(xi, yi):
                        keepf.append(d)
                        minXnew = min(minXnew, d[0])
```

```
                        minYnew = min(minYnew, d[1])
                        maxXnew = max(maxXnew, d[0])
                        maxYnew = max(maxYnew, d[1])
        dataf = keepf
        minX = minXnew
        maxX = maxXnew
        minY = minYnew
        maxY = maxYnew

        #new image for filtered detections
        W = int( ((maxX-minX)/cal.pixelWidth)+(2*pad) )
        H = int( ((maxY-minY)/cal.pixelHeight)+(2*pad) )
        fp = FloatProcessor(W, H)
        maxn = 0
        for p in dataf:
                xi = int( pad + ((p[0]-minX)/cal.pixelWidth) )
                yi = int( pad + ((p[1]-minY)/cal.pixelHeight) )
                value = fp.getf(xi,yi)+1
                maxn = max(value, maxn)
                fp.setf( xi, yi, value )

pores = getPoreRois(fp)

ol = Overlay()
nPores = 0
output = "Pore,X,Y,N Detections,"
for s in range(8):
        output += "Segment "+str(s)+","
output += "N Labelled,Labelling Efficiency\n"

headings = ["Pore","X","Y","Area (nm"+u'\u00B2'+")","Detections","
   Labelled Segments","Labelling Efficiency"]
for s in range(8):
        headings.append("Segment "+str(s+1))
results = []
```

```
for i,pore in enumerate(pores):
        IJ.showStatus("Pore "+str(i+1)+"/"+str(len(pores)))
        detections = []
        for p in dataf:
                xi = int( pad + ((p[0]-minX)/cal.pixelWidth) )
                yi = int( pad + ((p[1]-minY)/cal.pixelHeight) )
                if pore.contains(xi, yi):       #check using int
                    coords
                        detections.append( [xi, yi] )

        if len(detections) > 0:
                centroid = pore.getContourCentroid()    #px, image
                    coords
                segs = getSegCounts(detections, centroid)
                segAngles = segs[0]
                segCounts = segs[1]

                area = pore.getStatistics().area * cal.pixelWidth *
                    cal.pixelHeight

                nSegs = len( list(filter(lambda n:n>=minPpc,
                    segCounts)) )       #number of labelled segments
                resultLine = [nPores, centroid[0]*cal.pixelWidth,
                    centroid[1]*cal.pixelHeight, area, len(detections)
                    , nSegs, nSegs/8.0]

                pore.setStrokeColor(Color.CYAN)
                ol.add(pore)

                cx = centroid[0]
                cy = centroid[1]
                for i in range(len(segAngles)):
                        count = segCounts[i]
                        resultLine.append(count)
                        lineTheta1 = segAngles[i]
```

```
                    r = maths.sqrt(pore.getLength()/2*maths.pi)
                    lineX1 = cx + (maths.cos(lineTheta1)*r)
                    lineY1 = cy + (maths.sin(lineTheta1)*r)
                    line = Line(cx,cy, lineX1, lineY1)
                    line.setStrokeColor(Color.MAGENTA)
                    ol.add(line)

                    poly = FloatPolygon()
                    poly.addPoint(cx,cy)
                    poly.addPoint(lineX1,lineY1)
                    lineTheta2 = lineTheta1 + qpi

                    lineX2 = cx + (maths.cos(lineTheta2)*r)
                    lineY2 = cy + (maths.sin(lineTheta2)*r)
                    poly.addPoint(lineX2,lineY2)
                    polyRoi = PolygonRoi(poly, PolygonRoi.POLYGON
                        )

                    fillColour = Color(0,255,0,64) if segCounts[i
                        ] >= minPpc else Color(255,0,0,64)
                    polyRoi.setFillColor(fillColour)
                    ol.add(polyRoi)

                    #containedRoi = ShapeRoi(polyRoi).and(
                        ShapeRoi(pore))
                    #segArea = containedRoi.getStatistics().area
                        * cal.pixelWidth * cal.pixelHeight
                    #normCount = segCounts[i] / segArea      #
                        detections per mm^2
                    #resultLine.append(normCount)

                results.append(resultLine)
                nPores += 1

if showImage:
        imp = ImagePlus("Image maxn="+str(maxn), fp)
```

```
        imp.setDisplayRange(0, maxn)
        imp.setCalibration(cal)
        imp.setOverlay(ol)
        imp.show()

outPath = IJ.getFilePath("CSV Output...")
with open(outPath, 'w') as outFile:
        writer = csv.writer(outFile, lineterminator='\n')
        writer.writerow(headings)
        for line in results:
                writer.writerow(line)
```

Appendix D

Probabilistic model fit

Author: David Jordan

```
clear
clc
load('data.mat')
for i = 1:8
    h(i) = sum(data==i)/length(data);
end



dark = @(p) binopdf(0,4,p);
light = @(p) binopdf(1:8,8,1-dark(p));
binom_error = @(p) sum((h-light(p)).^2);

x0 = fminbnd(binom_error,0,1);

bar(h)
hold on
plot(light(x0),'r-o','linewidth',2)
title(['p_{label}=' num2str(x0)])
xlabel('Number of Corners')
ylabel('Fraction')
set(gca,'fontsize',14,'fontweight','bold')
legend({'data','fit'})
```

Appendix E

Single parameter fit

Author: David Jordan

```
function [lambda, error] = fitSingleParameter(data,t,p,nInts)

if length(p)==32
f = @(m,t,p) p(1)*normpdf(t,m,(1/sqrt(nInts))*(m))+...
    p(2)*normpdf(t,(m/2),(1/sqrt(nInts))*((m/2)))+...
    p(3)*normpdf(t,(m/3),(1/sqrt(nInts))*((m/3)))+...
    p(4)*normpdf(t,(m/4),(1/sqrt(nInts))*((m/4)))+...
    p(5)*normpdf(t,(m/5),(1/sqrt(nInts))*((m/5)))+...
    p(6)*normpdf(t,(m/6),(1/sqrt(nInts))*((m/6)))+...
    p(7)*normpdf(t,(m/7),(1/sqrt(nInts))*((m/7)))+...
    p(8)*normpdf(t,(m/8),(1/sqrt(nInts))*((m/8)))+...
    p(9)*normpdf(t,(m/9),(1/sqrt(nInts))*((m/9)))+...
    p(10)*normpdf(t,(m/10),(1/sqrt(nInts))*((m/10)))+...
    p(11)*normpdf(t,(m/11),(1/sqrt(nInts))*((m/11)))+...
    p(12)*normpdf(t,(m/12),(1/sqrt(nInts))*((m/12)))+...
    p(13)*normpdf(t,(m/13),(1/sqrt(nInts))*((m/13)))+...
    p(14)*normpdf(t,(m/14),(1/sqrt(nInts))*((m/14)))+...
    p(15)*normpdf(t,(m/15),(1/sqrt(nInts))*((m/15)))+...
    p(16)*normpdf(t,(m/16),(1/sqrt(nInts))*((m/16)))+...
    p(17)*normpdf(t,(m/17),(1/sqrt(nInts))*((m/17)))+...
    p(18)*normpdf(t,(m/18),(1/sqrt(nInts))*((m/18)))+...
    p(19)*normpdf(t,(m/19),(1/sqrt(nInts))*((m/19)))+...
    p(20)*normpdf(t,(m/20),(1/sqrt(nInts))*((m/20)))+...
```

```
    p(21)*normpdf(t,(m/21),(1/sqrt(nInts))*((m/21)))+...
    p(22)*normpdf(t,(m/22),(1/sqrt(nInts))*((m/22)))+...
    p(23)*normpdf(t,(m/23),(1/sqrt(nInts))*((m/23)))+...
    p(24)*normpdf(t,(m/24),(1/sqrt(nInts))*((m/24)))+...
    p(25)*normpdf(t,(m/25),(1/sqrt(nInts))*((m/25)))+...
    p(26)*normpdf(t,(m/26),(1/sqrt(nInts))*((m/26)))+...
    p(27)*normpdf(t,(m/27),(1/sqrt(nInts))*((m/27)))+...
    p(28)*normpdf(t,(m/28),(1/sqrt(nInts))*((m/28)))+...
    p(29)*normpdf(t,(m/29),(1/sqrt(nInts))*((m/29)))+...
    p(30)*normpdf(t,(m/30),(1/sqrt(nInts))*((m/30)))+...
    p(31)*normpdf(t,(m/31),(1/sqrt(nInts))*((m/31)))+...
    p(32)*normpdf(t,(m/32),(1/sqrt(nInts))*((m/32)));
elseif length(p)==4
   f = @(m,t,p) p(1)*normpdf(t,m,(1/sqrt(nInts))*(m))+...
    p(2)*normpdf(t,(m/2),(1/sqrt(nInts))*((m/2)))+...
    p(3)*normpdf(t,(m/3),(1/sqrt(nInts))*((m/3)))+...
    p(4)*normpdf(t,(m/4),(1/sqrt(nInts))*((m/4)));
else
    disp('Error')
end

y = ksdensity(data,t,'function','pdf');
[lambda,~] = fminsearch(@(b) sum((y - f(b,t,p)).^2), prctile(data,90)
   );

error = sum((y - f(lambda,t,p)).^2);
figure
h = histogram(data(:),t,'Normalization','pdf');
hold on
centers = cumsum(diff(h.BinEdges))-diff(h.BinEdges(1:2))/2;
plot(centers,f(lambda,centers,p),'-ro')
end
```

Appendix F

Two parameter fit

Author: David Jordan

```
function lambda = fitDoubleParameter(data,t,nInts)

y = ksdensity(data,t,'function','pdf');
[lambda,~] = fminsearch(@(b)
sum((y - twoParameterMixture(b,t,nInts)).^2), [300  0.2]);

figure
h = histogram(data(:),t,'Normalization','pdf');
hold on
centers = cumsum(diff(h.BinEdges))-diff(h.BinEdges(1:2))/2;
plot(centers,twoParameterMixture(lambda,centers,nInts),'-ro')
```

Appendix G

Pair correlation function fit

The first step in the analysis of protein clusters is to prove the existence of double clustering in the localisation data. It was discussed in Section 5.2.1 that molecular blinking pertaining to the same fluorescent protein is generating a small clustering structure (with radius 10 nm) in the data. In order to show the existence of a biological clustering (protein clusters) the approach described in (Sengupta et al. 2011) is adapted. The approach is based on a summary statistic, called *pair correlation function* computed from the localisation data. Summary statistic functions measure characteristics of spatial patterns and are multi-scale generalisations of the classical statistics (Illian et al. 2008).

Many summary statistics make use of the *intensity measure (average point density)* of a point process, given by:

$$\Lambda(B) = \mathbb{E}(N(B)) \tag{G.1}$$

where $N(B)$ is the random number of points in the set B. Under certain continuity conditions:

$$\Lambda(B) = \int_B \lambda(x) \mathrm{d}\, x, \tag{G.2}$$

where $\lambda(x)$ is called *intensity function*. It describes the probability that the point process contains a point at location x.

By generalisation, the *second order density* $\rho^{(2)}$ is the probability that at positions x and y there exist points of the point pattern: if $b(x)$ is the disk centered at point x of area $\mathrm{d}\, x$,

$d(x, y) = r$, the probability that $b(x)$ and $b(y)$ both contain a localisation is

$$[p_2(x, y) = \rho^{(2)}(x, y)\mathrm{d}\, x\mathrm{d}\, y \tag{G.3}$$

Finally, one can define the *pair correlation function* as:

$$[g(r) = \frac{\rho^{(2)}(r)}{\lambda^2} \tag{G.4}$$

Alternatively, given a typical point o, the probability that there exists another point of the process inside $b(x)$ is $\lambda \cdot g(r)\mathrm{d}\, x$.

It is known that:

- For a Poisson process (describing a complete spatial random process): $g(r) = 1$.
- For regular (inhibition) processes: $g(r) < 1$.
- For cluster processes: $g(r) > 1$.

By computing the empirical point correlation function of a process and comparing it to the constant function of value 1 (or the point correlation function of simulated completely spatial random process) it is possible to draw conclusions with respect to the process in question being clustered or not.

Moreover the analytical expression for the Thomas process is known to be:

$$g(r) = 1 + exp(-r^2/(4\sigma^2))/(4\pi\kappa\sigma^2) \tag{G.5}$$

where $\lambda = \kappa * \mu$ is the intensity of the process and mu is the expected number of points in the cluster). By fitting (e.g. *mincontrast*) the analytical expression to the computed empirical correlation function, one can obtain estimates of average cluster radius and average number of points inside the clusters.

In a pair-correlation analysis, pc-PALM, is proposed order to analyze patterns of protein organization across the plasma membrane in COS-7. The two model compared were a

simple single cluster model and a more complex double cluster model:

$$g_1(r) = g(r)^{stoch} + 1 \tag{G.6}$$

$$\begin{aligned} g_2(r) &= g(r)^{stoch} + \\ &\quad + (A\exp(-r/\xi) + 1) * g(r)^{psf} \end{aligned} \tag{G.7}$$

The function giving a better goodness of fit to protein localisation data was the double cluster structure.

In (Hwang et al. 2019) g_2 is replaced with a more intuitive double Thomas process (both levels of clusters are 2d Gaussian distributed around thecenters), with pair correlation function given by:

$$g_2(r) = 1 + \frac{1}{(4\pi\kappa_1\sigma^2)} exp\left(\frac{-r^2}{4\sigma^2}\right) + \frac{1}{(4\pi\kappa_2(\sigma_2^2+\sigma^2)} \exp\left(\frac{-r^2}{4(\sigma_2^2+\sigma^2)}\right). \tag{G.8}$$

This case corresponds to the protein cluster model described in the Section 5.2.1, with $sigma$ the protein cluster radius and $sigma_2$ the scale localisation precision corresponding to the cluster of blinks of the same fluorescent protein.

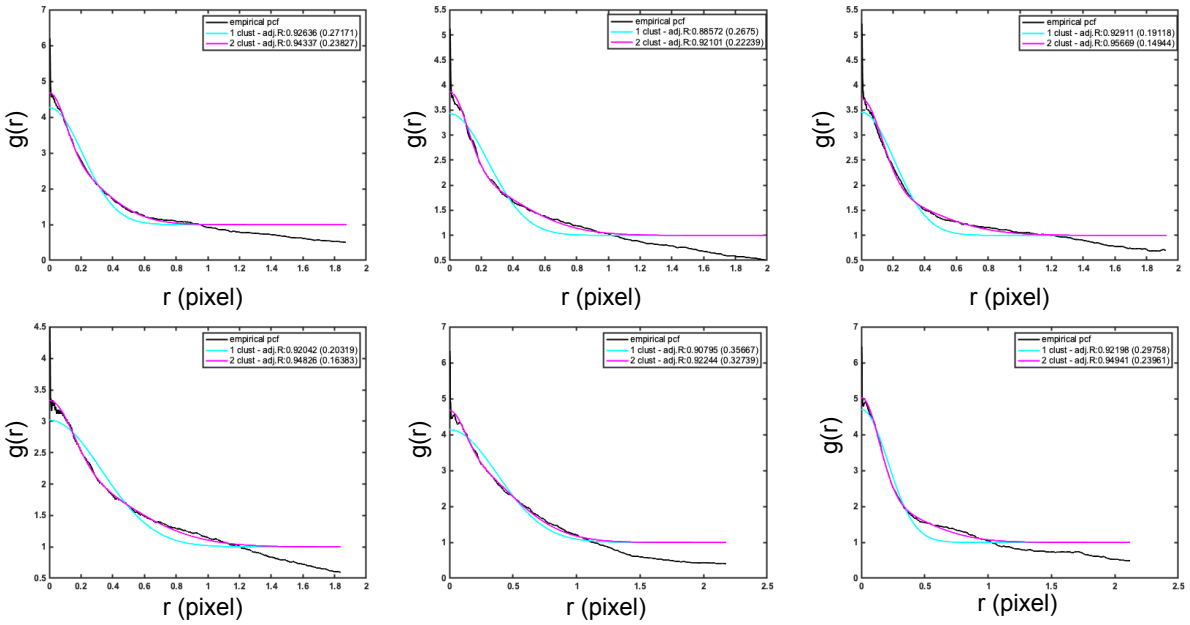


Figure G.1 **Validation of real clustering of aPKC-Halo by PCF analysis.** A PCF plot for aPKC-Halo along a junction in follicle epithelial cells.

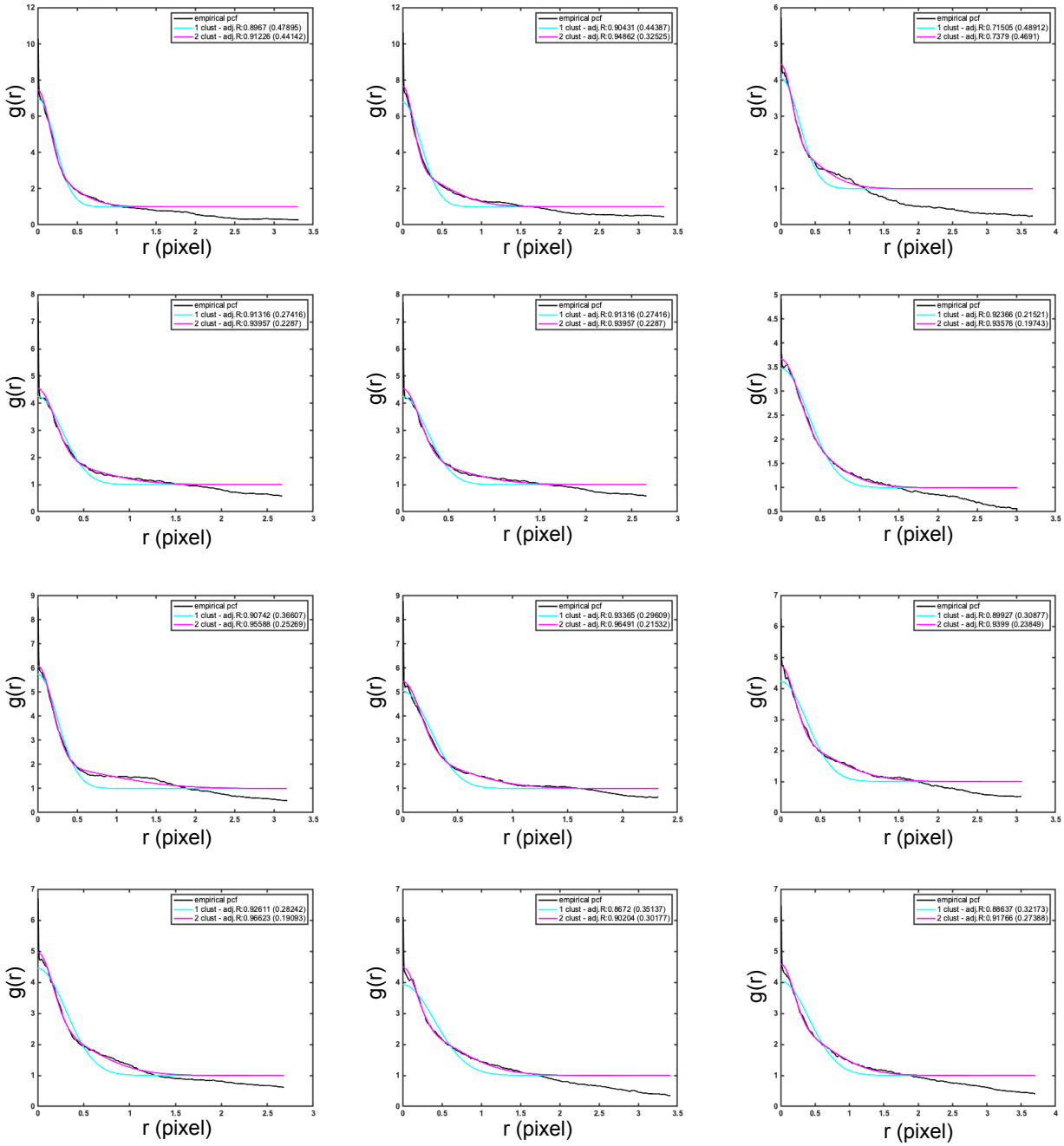


Figure G.2 **Validation of real clustering of Crumbs-Halo by PCF analysis.** A PCF plot for Crumbs-Halo along a junction in follicle epithelial cells.

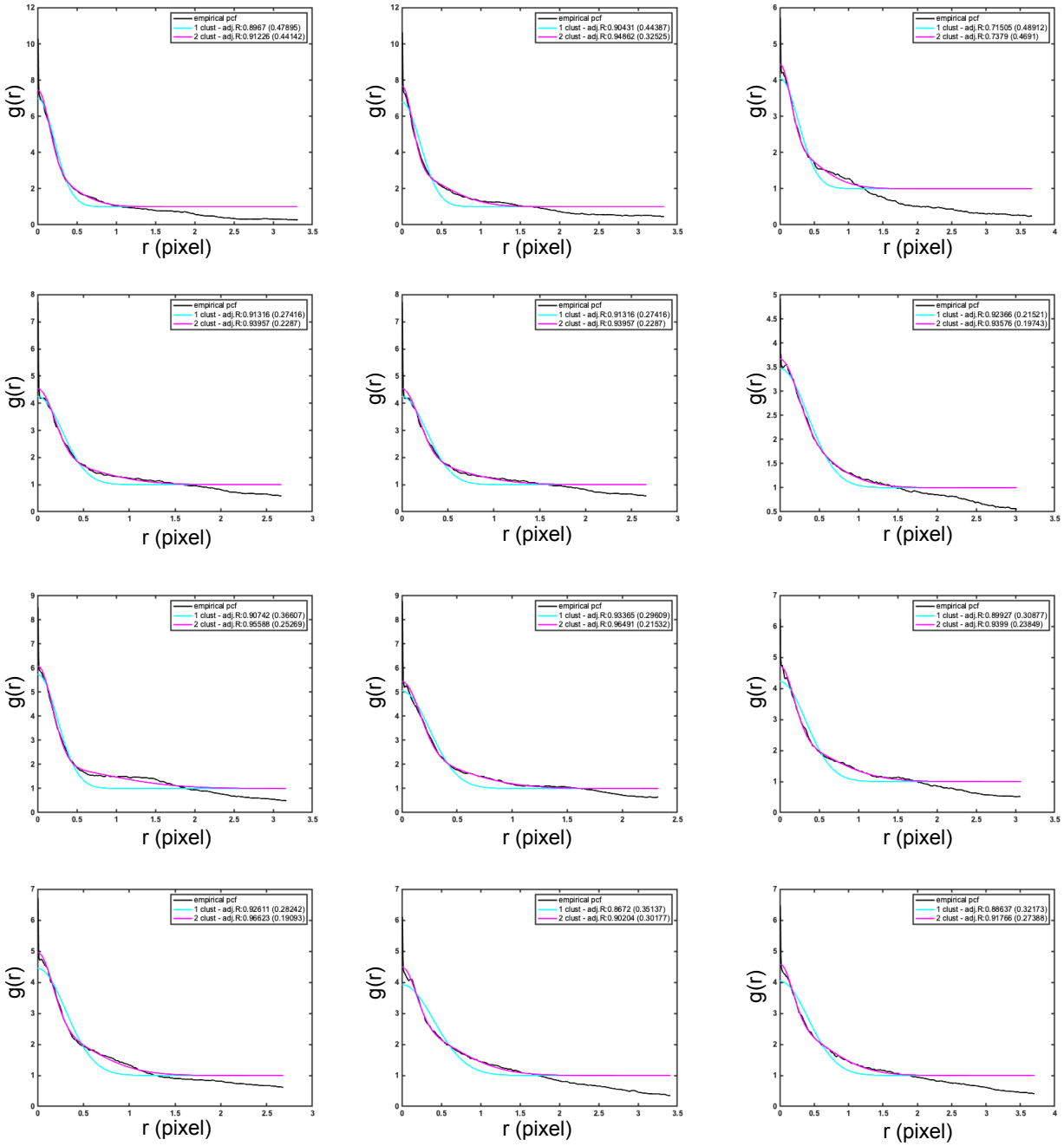


Figure G.3 **Validation of real clustering of Par6-Halo by PCF analysis.** A PCF plot for Par6-Halo along a junction in follicle epithelial cells.

Appendix H

Cluster analysis

For using mean shift in Matlab *pdollar* toolbox was used: https://pdollar.github.io/toolbox/
For using yaml files in Matlab *+yaml* toolbox was used: https://code.google.com/p/yamlmatlab/

Author: Leila Muresan

```
addpath .\pdollar\classify % meanshift
addpath '.' % read yaml, PoissonAnalysisFc
% Parameters:
influx = 0.0005387
label_eff = 1 % = 1 means 100%
qP = 1 % = 1 qPaint estimates nr. of molecules in the whole ROI, 0 -
    performs meanshift + qPaint (mol/cluster)
no_clean = 0 % if 1 it does not clean the data before qPaint
png  = 1; % = 1 saves figure as png, = 0 saves as fig
nradius = 1; % bandwidth for meanshift
resdir = uigetdir('','Select result folder'); % Results folder

%% Select Roi file
addpath C:\Leila\yamlmatlab-master

[fy dy] = uigetfile('*.yaml');
yaml_file = fullfile(dy,fy);
ry = yaml.ReadYaml(yaml_file);
circles =  ~isempty(strfind(ry.Shape , 'Circle'))
```

```
clear rct
if ~circles % rectangle ROIs
    ROIc = ry.Center0x2DAxis0x2DPoints;%csvread('C:\Leila\
        aPKC_data_Leila\20190621\RoisArea2.csv')
    width = ry.Width/2;
    nrroi  = 1;
    for k = 1:length(ROIc)

        ROI = cell2mat(ROIc{k});
        %ROI = [ROIc{k};ROIc{k}]+1.6*[-1 1; -1 1]
        m = (ROI(2,2)-ROI(1,2))/(ROI(2,1)-ROI(1,1));
        m2 = (1/m^2+1);
        if (m~=1)
            rct{nrroi} = -[roots([1 2*ROI(1,1) ROI(1,1)^2-(width^2)/
                m2]); roots([1 2*ROI(2,1) ROI(2,1)^2-(width^2)/m2])];
            rct{nrroi} = cat(2, rct{nrroi}, [ROI(1,2)-1/m*(rct{nrroi
                }(1)-ROI(1,1)); ROI(1,2)-1/m*(rct{nrroi}(2)-ROI(1,1));
                 ROI(2,2)-1/m*(rct{nrroi}(3)-ROI(2,1)); ROI(2,2)-1/m*(
                rct{nrroi}(4)-ROI(2,1))]);
            nrroi = nrroi+1;
        end
    end
    %figure; plot((rct{1}([1 3 4 2]',1))',(rct{1}([1 3 4 2]',2))')
else %  circles ROIs
    ROIc = ry.Centers;%csvread('C:\Leila\aPKC_data_Leila\20190621\
        RoisArea2.csv')
    radc = ry.Diameter/2;
    nrroi  = 1;
    for k = 1:length(ROIc)
        % TODO: make proper circles!
        ROI = [ROIc{k};ROIc{k}]+radc*[-1 1; -1 1]
        rct{nrroi} =  repmat([ROIc{k,1} ROIc{k,2} ],[4 1])+0.8*[-1
            -1; -1 +1; 1 -1; 1 1];
        nrroi = nrroi+1;

    end
```

```
    figure; plot((rct{5}([1 3 4 2]',1))',(rct{5}([1 3 4 2]',2))')
end
%% read mat-file data
[fname dname] = uigetfile('*.mat');
load(fullfile(dname, fname));
fn = strtok(strrep(dname, '\', '_'), '.')
fn = strrep(fn,':', '')
npts = [handles.reconstruction(:,1:3)];
%npts = [handles.reconstruction(:,1:3); handles.reject_mol(:,1:3)];
figure ; plot(npts(:,1), npts(:,2), 'm.', 'MarkerSize', 0.5)
axis equal tight%%
imsize = ceil([max(npts(:,1)) max(npts(:,2))]);
npts(:,1:2) = max(npts(:,1:2), [1 1]);


%% Analysis
clear res pos BIC T

figure(1);
clf
for sel = 1:length(rct)
    imsize = [floor(max(npts(:,2)))+1, floor(max(npts(:,1)))+1];
    figure(1);
    h_im =imshow(zeros(imsize(1), imsize(2))); hold on
    plot(npts(:,1),npts(:,2), 'm.', 'MarkerSize', 0.5);
    %e = impoly();
    e = impoly(gca,rct{sel}([1 3 4 2 ]',:),'Closed',true);
    % h = impoly(gca,[188,30; 189,142; 93,141; 13,41; 14,29]);
    BW = createMask(e,h_im);
    %mwrite(uint8(BW),strcat(folder, strrep(files(k).name, '.csv',
        strcat('maskForGMM',num2str(n), '.tif'))), 'tiff');
    id = find(BW(sub2ind(size(BW), floor(npts(:,2)),floor(npts(:,1)))
        )>0);
    plot(npts(id,1),npts(id,2), 'c.'); hold on
    res{sel} = [ 0 0 0 0 0 0 0]; BIC{sel} = [0 0 0 0 0 0]; T{sel} =
        [0 0 0 0];
```

```
    if qP
        if no_clean
            [res{sel}  Ttau2(sel) Ttau(sel) muhat(sel)] = qPaint(npts
                (id,:),influx);
        else
            [EM MM delta]= PoissonAnalysisFc(npts(id,:), max(npts
                (:,1:2)), 10, 0, 1,  '');
            ii = find(delta>0);
            res{sel} = qPaint(npts(id(ii),:),influx);
        end
        text(rct{sel}(1,1),rct{sel}(1,2),num2str(res{sel}(end)));
%         csvwrite( strcat(resdir, fn, 'Sel',num2str(sel),'.csv'),
    npts(id,:));
%         csvwrite(strcat(resdir, fn, 'Win',num2str(sel),'.csv') ,
    rct{sel}([1 3 4 2 ]',:));

    else
        % clean data if less than 10 blinks
        [EM MM delta]= PoissonAnalysisFc(npts(id,:), max(npts(:,1:2))
            , 10, 0, 1,  '');
        ii = find(delta>0);
        if length(ii)<2
            res{sel} = [ 0 0 0 0 0 0 0]; BIC{sel} = [0 0 0 0 0 0]; T{
                sel} = [0 0 0 0];
        else
            [res{sel}  pos{sel} BIC{sel} T{sel} ]= AnalyseDataShort(
                npts(id(ii),:), nradius, fullfile(resdir,strcat(strtok
                (fy, '.'), 'Res',num2str(sel))), 1, influx, png);
        end
        if ~isempty(T{sel})
            text(rct{sel}(1,1),rct{sel}(1,2),num2str(mean(T{sel}(:,
                end))));
        else
            text(rct{sel}(1,1),rct{sel}(1,2),'0');
        end
%         csvwrite( strcat(resdir, fy, 'Sel',num2str(sel),'.csv'),
```

```
    npts(id,:));
%           csvwrite(strcat(resdir, fy, 'Win',num2str(sel),'.csv') ,
    rct{sel}([1 3 4 2 ]',:));

     end
     pause(1)
end
close all
```

where function **qPaint** is

```
function [T Ttau2 Ttau muhat]= qPaint(x,flux);%, radius, radius2,
    nrComp, fast, n, png, withkde)
%       addpath C:\Leila\meanShift
%       addpath C:\Leila\Simulations
     T= [0 0 0 0];

     %%ii = find(IDX == uclust(i));
     plot(x(:,1),x(:,2), '.'); hold on
     axis equal tight
     t = sort(x(:,end));
     %Ttau2 = t;
     Tseennr = 1;
     Testnr = 1;
     Testnr2 = 1;
     Testnr3 =1;
     Testnr4 =1;
     if length(t)>2
          dt = diff(t);
          offtimes = abs(dt(find(dt~=1)));
          if length(offtimes)>2
          length(offtimes)
          %p(unique(x(:, 4)),4)
          [muhat,muci] = expfit(offtimes);

               [f,tpoints,flo,fup] = ecdf(offtimes);
               fo = fitoptions('Method','NonlinearLeastSquares',...
```

```
                'Lower',[0 0 0],...
                'Upper',[Inf Inf Inf],...
                'StartPoint',[muhat 0 1 ]);
ft = fittype('c*(1-exp(-x/a))+b','independent','x', '
    coefficients',{'a', 'b','c'},'options',fo);
Ttau = interp1(f, tpoints, 1-1/exp(1));
if length(f)<3
    mu2.a = muhat;
else
    [mu2,gof] = fit(tpoints,f,ft);
end
max(tpoints)
if max(tpoints)==0
    offtimes
    figure; plot(t)
    %Tseennr = sum(p(unique(x(:, 4)),4));
    Testnr =5000;
    Testnr2 =5000;
    Testnr3 =5000;
    Testnr4 =5000;
else
tt = 0:max(tpoints)/100:max(tpoints)
valtt = 1-exp(-1/mu2.a*tt);
if length(valtt)~=length(unique(valtt))
    idx = find(1-1/exp(1)>1-exp(-1/mu2.a*tt));
    Ttau2 = tt(idx(1));
else
    Ttau2 = interp1( 1-exp(-1/mu2.a*tt), tt, 1-1/exp(1));
end
%Ttau2 = interp1( 1-exp(-1/mu2.a*tt), tt, 1-1/exp(1));
%Tseennr = sum(p(unique(x(:, 4)),4));
T(1) =1/(flux*muhat);
T(2) =1/(flux*mu2.a) ;
T(3) =1/(flux*Ttau);
%Testnr4 =1/(length(find(diff(x(:,6))~=1))/(10000*length(
    p(:,4)))*(Ttau2));
```

```
            T(4) = 1/(flux*Ttau2);

            end
        end
    end
end

% figure;
% plot(Testnr3, 'rx');hold on ;
% %plot(Tseennr, 'bo','LineWidth',2); hold on
% plot(Testnr, 'go');
% %plot(Testnr2, 'mx');
% plot(Testnr4, 'cx', 'LineWidth',2);
% xlabel('Cluster');
% ylabel('Number of proteins');
% legend('Estimated')
%
%
```

where function **PoissonAnalysisFc** is

```
%% SpatPattern/NNRaftery.pdf
% Leila Muresan, 2007
% centroids - 2dim vect. with positions
% sz - size of the window/image
% kNN - the order of the max NN distance
% imageon - with or wothout images
% v = 0 - sampled version,
%   = 1 - full version of the distance matrix
% fname - file with results


function [EMMM delta]= PoissonAnalysisFc(centroids, sz, kNN, imageon
    , v, fname)
%addpath 'D:\Leila\src\TautString\'
% sz - size of the image
if ~exist('kNN')
```

```
    kNN =  50;
end
format long
if ~exist('imageon')
    imageon = 0;
end
if ~exist('v')
    v = 1;
end
%% Mixture based on NN
kNN = min(size(centroids,1)-1, kNN);

rMom = [];
for ds = 10:10:100

    for i = 1:sz(1)/ds
        for j = 1:sz(2)/ds
            Summary(i,j) = size(find((centroids(:,1) <= j*ds) & (
                centroids(:,2) <= i*ds) & (centroids(:,1) >(j-1)*ds) &
                 (centroids(:,2) > (i-1)*ds)),1);
        end
    end


    %% Analysis . Method of Moments (See Fruhwirth)
    % Implement with 3 classes!

    Counts = Summary(:);
    %added this
    Counts(Counts==0)=[];
    %R =reshape(Counts, size(Summary));

    mom1 = mean(Counts);
    mom2 = 1/length(Counts)*sum( Counts.* (Counts-1) );
    mom3 = 1/length(Counts)*sum( Counts.* (Counts-1).*(Counts-2) );
```

```
    if (mom2- mom1^2 ~= 0)
        b = (mom3 - mom1*mom2)/(mom2- mom1^2);
    else
        b = 0;
    end
    c = mom1*b-mom2;

    mu1 = (b+sqrt(b^2-4*c))/2;
    mu2 = (b-sqrt(b^2-4*c))/2;

    if mu2~=mu1
        eta1 = (mu2-mom1)/(mu2-mu1);
    else
        eta1 = 0;
    end
    if mu1 > mu2
        if (eta1*mu1+(1-eta1)*mu2) ~= 0
            pi1 = (eta1*mu1)/(eta1*mu1+(1-eta1)*mu2);
        else pi1 = 0;
        end

    else

        if (eta1*mu1+(1-eta1)*mu2) ~= 0
            pi1 = (1-eta1)*mu2/(eta1*mu1+(1-eta1)*mu2);
        else pi1 = 0;
        end
    end


    rMom = [rMom; max(mu1/(ds^2), mu2/(ds^2)), min(mu1/(ds^2), mu2/(
        ds^2)), pi1];

end

% if imageon
```

```
%       figure; plot( rMom(:,1), 'b*'); hold on
%       plot( rMom(:,2), 'o', 'Color',[.6 .6 .6]); hold on
% end
MM = rMom(2,:);
clear Summary
%% Raftery
EmMethod = 1;
EM = [0, 0, 0, 0 ];
delta = 0;
if EmMethod
% correct for edge effects
size(centroids)

% distance  computation
%d = distMatrix (centroids, centroids);

%
%d = distMatrixS (centroids, kNN);
id = 1:size(centroids,1);
if v == 1
    d = distMatrixSTorus (centroids, kNN, sz(2), sz(1));
else
    [d id] = distMatrixSTorusSampled (centroids, kNN, sz(2), sz(1),
       1000);
end

[B,IX] = sort(d,2);

% EM for the model: d ~ p*Gamma1/2(K, l1*pi)+(1-p)*Gamma1/2(K, l2*pi)
delta = zeros(size(centroids,1),1);
es =[];
e2 = [];
res = [];
for k = 3:kNN;

   p = MM(3);
```

```
    lambda1 = MM(1);% /(20^2); %k/(pi*sum(B(:,k).^2));

    lambda2 = MM(2);% /(20^2); %k/(pi*sum(B(:,k).^2));

%       lambda1 = 0.05
%       lambda2 = 0.01
        cont = 1;
        pold = 5;
        eps = 0.001;
        col1 = [];
        col2 = [];




        while cont
        %for i = 1:30
            % E-step
            delta = p*transfGamma(B(:,k-1),lambda1*pi, k)./(p*transfGamma
                (B(:,k-1),lambda1*pi, k)+(1-p)*transfGamma(B(:,k-1),
                lambda2*pi, k));
            dd = delta;

                delta = round(delta);  % transform to 0 or 1

            % M-step
            lambda1 = double(k*sum(delta))/ max(eps,(pi*sum(B(:,k-1).^2.*
                delta)));
            lambda2 = double(k*sum(1-delta))/ max(eps,(pi*sum(B(:,k-1)
                .^2.*(1-delta))));
            p = sum(delta)/length(delta);
           % if (abs(p-pold)/pold < eps) & (abs(lambda1-lambda1old)/
              lambda1old < eps)&(abs(lambda2-lambda2old)/lambda2old < eps
              ) cont = 0; end
           cont = cont+1;
           if (cont>100)|((abs(p-pold) < eps) & (abs(lambda1-lambda1old)
```

```
        < eps)&(abs(lambda2-lambda2old) < eps)) cont = 0; end

        pold = p;
        lambda2old =lambda2;
        lambda1old =lambda1;
        col1 = [col1 lambda1];
        col2 = [col2 lambda2];
    end

    if imageon
        figure;
        subplot(211)
        [n,x] = hist(B(:,k-1), 200);
        hist(B(:,k-1), 200); title(num2str(k))
        subplot(212)
        y= p*transfGamma(x, lambda1*pi, k)+(1-p)*transfGamma(x,
            lambda2*pi, k);
        %bar(x, n/sum(n), 'b')
        hold on
        plot(x,y, 'r*-')
    end

    %%Automatic selection of the kNN via peicewise constant
        approximation

    % I CHANGED HERE - Feb09 es with e2
    es = [es; -entropy(real(dd))];
    e2= [e2; -entropy(real(delta))];


%       lambda1
%       lambda2
%       p
%
%       mu1/10000
%       mu2/10000
```

```
%
%      eta1

    res = [res; lambda1  lambda2 p -entropy(real(dd)) ];

    %% plot: result of detection
    if imageon

%         figure(10);
%         idx1 = find (delta == 1);
%         plot(centroids(idx1,1), centroids(idx1,2), 'k*'); hold on;
   axis ij; %title('Signal separation based on EM algorithm');
%         idx2 = find (delta == 0);
%         plot(centroids(idx2,1), centroids(idx2,2), 'o', 'Color',[.3
    .3 .3]); hold on; axis ij ;
%         legend('Signal', 'Background');
%         axis equal tight
%         pause;
%         close all
        figure(10);
        idx1 = id(find (delta == 1));
        plot(centroids(idx1,1), centroids(idx1,2), 'k*'); hold on;
           axis ij; %title('Signal separation based on EM algorithm')
           ;
        idx2 = id(find (delta == 0));
        plot(centroids(idx2,1), centroids(idx2,2), 'o', 'Color',[.3
           .3 .3]); hold on; axis ij ;
        legend('Signal', 'Background');
        axis equal tight
        pause;
        close all
    end
end
if imageon
    figure;
    plot(es, 'r-*'); hold on
```

```
    plot(e2, 'b-+');
    pause
end
id1=kNN-1;
id2=kNN-1;
%% [f1, f2] = TautStringSmoothS(1:length(es), es, 0.5);
%% id1 = find(diff(f1)>0);
%% id2 = find(diff(f2)>0);
%% sel1 = kNN-1;
%% sel2 = kNN-1;
%%
%% if length(id1)>0 sel1 = min(kNN, id1(end)+1); end;
%% if length(id2)>0 sel2 = min(kNN, id2(end)+1); end;
%%% col1 = zeros(size(res,1),1);
%%% col1(sel1) = 1;
%%%
%%% col2 = zeros(size(res,1),1);
%%% col2(sel2) = 1;
%%%res = [res col1 col2 ];
%%
%% if sel2>0
%%     EM = res(sel2,:)
%%     if sum(isnan(EM)) >0
%%         EM = [0, 0, 0, 0 ];
%%     end
%%     MM
%%     cont = 1;
%%     delta = round( p*transfGamma(B(:,sel2),lambda1*pi, sel2+1)./(
    p*transfGamma(B(:,sel2),lambda1*pi, sel2+1)+(1-p)*transfGamma(B(:,
    sel2),lambda2*pi, sel2+1)));
%% else
%%     EM = [0, 0, 0, 0 ];
%%     cont = 0;
%% end
%%
%         figure;
```

```
%          %imagesc(A); title('Result')
%          %subplot(121);
%          hist(B(:,i), 200); %title('k-th nearest neighbour distances
    ');
%          h = findobj(gca,'Type','patch');
%          set(h,'FaceColor',[0.2 0.2 0.2],'EdgeColor','w')
%          %subplot(122);

end
if imageon

    figure(10);
    id = 1:length(delta);
    idx1 = id(find (delta == 1));
    plot(centroids(idx1,1), centroids(idx1,2), 'k*'); hold on; axis
        ij; %title('Signal separation based on EM algorithm');
    idx2 = id(find (delta == 0));
    plot(centroids(idx2,1), centroids(idx2,2), 'o', 'Color',[.6 .6
        .6]); hold on; axis ij ;
    legend('Signal', 'Background');
    axis equal tight
    saveIm = frame2im(getframe(10));
    % save image

    %imwrite(saveIm, strcat(fname,'SB.tif'), 'TIF');

%     figure;
%    plot(es, 'b*-'); hold on
%     plot(f1, 'r*-')
%     plot(f2, 'g*-')
%     plot(e2, 'r*-')
%     title('Entropies')
%     pause
%     close all
end
%% Save results
```

```
%csvwrite( fname, res)

%%

function y = transfGamma(x, L, K)

%y = exp(-L*pi*x^2)*2*(L*pi)^K*x^(2*K-1)/(K-1)!

if L>0
    % do i need 1/2 in front for sqrt
    %logv = 1/2*( -L*x.^2+log(2)+ K*log(L)+(2*K-1)*log(x) - sum(log
        (1:K-1)));
    logv = ( -L*x.^2+log(2)+ K*log(L)+(2*K-1)*log(x) - sum(log(1:K-1)
        ));
    maxv = max(logv);
    %y = sqrt(exp(logv-maxv)*exp(maxv));
    y = exp(logv-maxv)*exp(maxv);
else
    y= zeros(size(x));
end
```

where function **AnalyseDataShort** is

```
% nradius - bandwidth for meanshift
function [res  pos BIC Testnr ]= AnalyseDataShort(x,nradius, resdir,
    nFrames, flux,png);%, radius, radius2, nrComp, fast, n, png,
    withkde)
     addpath C:\Leila\meanShift
     addpath C:\Leila\Simulations

% simple kde estimate
[bandwidth,density,X,Y]=kde2d(x(:,1:2));
if (~exist('nradius'))
    nradius = max(bandwidth);
    nradius =0.5;
end
```

```
%https://xcorr.net/2015/04/06/calling-r-from-matlab-flat-file-
    communication/
%!R.exe BATCH C:/Leila/MatlabNestedClusterFit.R

%RunRcode('C:\Leila\Analysis\MatlabNestedClusterFit.R')
%nradius = max(nradius,1);

% [indices, IDX, C] = dbscan(x(:,1:2)', nradius, 20 );
% C = C';
% tic

% [clustCent,point2cluster,clustMembsCell] = MeanShiftCluster(x
    (:,1:2)', nradius);
% toc

[IDXms,Cms] = meanShift(x(:,1:2), nradius, .2, 100 ,9);%, 100 );
% Jungmann
uclust = unique(IDXms);
uclust(uclust ==-1) = [];

%flux = 0.0003215; in frames 1/(length(find(diff(x(:,6))~=1))/(
    nFrames*sum(p(:,4)))
clear Tseennr Testnr Testnr muhat mu2
h = figure
subplot(121)
Testnr = [];
for m = 1:length(uclust)
    ii = find(IDXms == uclust(m));
    %%ii = find(IDX == uclust(i));
    plot(x(ii,1),x(ii,2), '.','MarkerSize',2); hold on
    axis equal tight
    t = sort(x(ii,end));
    %Ttau2(m) = t;
    Testnr(m,1) = 1;
    Testnr(m,2) = 1;
    Testnr(m,3) = 1;
```

```
Testnr(m,4) =1;

if length(t)>2
    dt = diff(t);
    offtimes = abs(dt(find(dt~=1)));
    if length(offtimes)>2
    length(offtimes);
    %p(unique(x(ii, 4)),4)
    [muhat(m),muci] = expfit(offtimes);

        [f,tpoints,flo,fup] = ecdf(offtimes);
        fo = fitoptions('Method','NonlinearLeastSquares',...
                        'Lower',[0 0 0],...
                        'Upper',[Inf Inf Inf],...
                        'StartPoint',[muhat(m) 0 1 ]);
        ft = fittype('c*(1-exp(-x/a))+b','independent','x', '
           coefficients',{'a', 'b','c'},'options',fo);
        Ttau(m) = interp1(f, tpoints, 1-1/exp(1));
        if length(f)<3
            mu2.a = muhat(m);
        else
            [mu2,gof] = fit(tpoints,f,ft);
        end

        if max(tpoints)==0
            offtimes
            figure; plot(t)
            %Tseennr(m) = sum(p(unique(x(ii, 4)),4));
            Testnr(m,1) =5000;
            Testnr(m,2) =5000;
            Testnr(m,3) =5000;
            Testnr(m,4) =5000;
        else
        tt = 0:max(tpoints)/100:max(tpoints);
        valtt = 1-exp(-1/mu2.a*tt);
        if length(valtt)~=length(unique(valtt))
```

```
                idx = find(1-1/exp(1)>1-exp(-1/mu2.a*tt));
                Ttau2(m) = tt(idx(1));
            else
                Ttau2(m) = interp1( 1-exp(-1/mu2.a*tt), tt, 1-1/exp
                    (1));
            end
            %Ttau2(m) = interp1( 1-exp(-1/mu2.a*tt), tt, 1-1/exp(1));
            %Tseennr(m) = sum(p(unique(x(ii, 4)),4));
            Testnr(m,1) =1/(flux*muhat(m));
            Testnr(m,2) =1/(flux*mu2.a)    ;
            Testnr(m,3) =1/(flux*Ttau(m));
            %Testnr4(m) =1/(length(find(diff(x(:,6))~=1))/(10000*
                length(p(:,4)))*(Ttau2(m)));
            Testnr(m,4) = 1/(flux*Ttau2(m));
            if isnan( Testnr(m,4)) || isinf( Testnr(m,4))
                Testnr(m,4) = Testnr(m,2);
            end
        end
    end
    end

end


% %plot(Tseennr, 'bo','LineWidth',2); hold on
% plot(Testnr, 'go');
% %plot(Testnr2, 'mx');
% plot(Testnr4, 'cx', 'LineWidth',2);
% xlabel('Cluster');
% ylabel('Number of proteins');
% legend('Estimated')



subplot(122)
plot(x(:,1), x(:,2), 'k.', 'LineWidth',2, 'MarkerSize', 2); hold on;
```

```
    title('Top view'); hold on
%plot(sc(1)*rand(300,1), sc(2)*rand(300,1),'k.', 'LineWidth',2,'
    MarkerSize', 2);% sc(3)*rand(n,1)];
axis tight equal
%figure; plot(C(:,1), C(:,2), '.'); hold on
    if ~png
        saveas(gcf, strcat(resdir,'Est.fig'), 'fig')
    else
        print(gcf,strcat(resdir,'Est.png'), '-dpng')
    end
plot(Cms(:,1), Cms(:,2), 'm.'); hold on
%estdbkde = cellfun('length', indices);
%estmskde = cellfun('length', indices)/10;

if ~isempty(Testnr)
    figure;
     plot(Testnr);hold on ;
    [pos BIC] = ModelComparison(Testnr(:,4));
else
    pos = []; BIC = [];
end

IDXms(IDXms ==-1)=[];
n = hist(IDXms, unique(IDXms));
if ~isempty(n)
    res = [ nradius mean(n)  var(n)  var(n)/mean(n) length(n) mean(
        Testnr(:,4)) std(Testnr(:,4))];
else
    res = zeros(1,7);
end
%res{3} =[nradius mean(estdbkde)  var(estdbkde)  var(estdbkde)/mean(
    estdbkde) length(estdbkde)];
%n = n/10;

%figure; hist(IDXms, unique(IDXms))
%[estx, estf, dbeflo,dbefup] = ecdf(cellfun('length', indices)/10);
```

```
%save(fullfile(resdir, strcat('DBSCANkdeCDFs', num2str(n),'_P.mat')),
      'estx', 'estf', 'estdbkde');
close all
```

where function **distMatrixSTorus** is

```
% Make distance matrix (small)
% input: two matrices. Each matrix consists of pairs
% of coordinates [x y]
% The coordinates are considered on a torus.
% Output: in each line the first k nearest neighbors of points in A
%
function [res id] = distMatrixSTorus (A, k, sx, sy);
% B=A;
% Dkx = A(:,1)*ones(1,size(B, 1));
% Dky = A(:,2)*ones(1,size(B, 1));
% Dsucckx = ( B(:,1)*ones(1,size(A, 1)) )';
% Dsuccky = ( B(:,2)*ones(1,size(A, 1)) )';
% res1 = sqrt((Dkx - Dsucckx).^2 + (Dky - Dsuccky).^2);

res = [];
for i = 1: size(A,1)
    %maxd = max(sqrt((A(:,1) - A(i,1)).^2 + (A(:,2) - A(i,2)).^2));
    % not really correct. I take the closest from the real point and
       it' 2
    % "toroidal" images
    s =sort( min( [sqrt((A(:,1) - A(i,1)).^2 + (A(:,2) - A(i,2)).^2)
       '; sqrt((sx +A(:,1) - A(i,1) ).^2 + (A(:,2) - A(i,2)).^2) ';...
        sqrt((A(:,1) - A(i,1) ).^2 + (sy + A(:,2) - A(i,2)).^2)';
           sqrt((-A(:,1)+1 - A(i,1) ).^2 + ( A(:,2) - A(i,2)).^2)';
           ...
        sqrt((-A(:,1)+1 - A(i,1) ).^2 + ( -A(:,2) - A(i,2)).^2)'] ));
%    s = sort([sqrt((A(:,1) - A(i,1)).^2 + (A(:,2) - A(i,2)).^2);
   sqrt((A(:,1) - A(i,1) - sx).^2 + (A(:,2) - A(i,2) -sy).^2) ]);
```

```
    res = [res; s(2:k+1)];
end

%sum(sum(sort(res1(:, 2:k+1),2)-res))


%figure; imagesc(DistM)
```

where function **kde2d** is

```
function [bandwidth,density,X,Y]=kde2d(data,n,MIN_XY,MAX_XY)
% fast and accurate state-of-the-art
% bivariate kernel density estimator
% with diagonal bandwidth matrix.
% The kernel is assumed to be Gaussian.
% The two bandwidth parameters are
% chosen optimally without ever
% using/assuming a parametric model for the data or any "rules of
    thumb".
% Unlike many other procedures, this one
% is immune to accuracy failures in the estimation of
% multimodal densities with widely separated modes (see examples).
% INPUTS: data - an N by 2 array with continuous data
%             n - size of the n by n grid over which the density is
    computed
%                 n has to be a power of 2, otherwise n=2^ceil(log2(n)
    );
%                 the default value is 2^8;
% MIN_XY,MAX_XY- limits of the bounding box over which the density is
     computed;
%                 the format is:
%                 MIN_XY=[lower_Xlim,lower_Ylim]
%                 MAX_XY=[upper_Xlim,upper_Ylim].
%                 The dafault limits are computed as:
%                 MAX=max(data,[],1); MIN=min(data,[],1); Range=MAX-
    MIN;
%                 MAX_XY=MAX+Range/4; MIN_XY=MIN-Range/4;
```

```
% OUTPUT: bandwidth - a row vector with the two optimal
%                    bandwidths for a bivaroate Gaussian kernel;
%                    the format is:
%                    bandwidth=[bandwidth_X, bandwidth_Y];
%          density  - an n by n matrix containing the density values
   over the n by n grid;
%                    density is not computed unless the function is
   asked for such an output;
%              X,Y  - the meshgrid over which the variable "density"
   has been computed;
%                    the intended usage is as follows:
%                    surf(X,Y,density)
% Example (simple Gaussian mixture)
% clear all
%  % generate a Gaussian mixture with distant modes
%   data=[randn(500,2);
%       randn(500,1)+3.5, randn(500,1);];
%  % call the routine
%     [bandwidth,density,X,Y]=kde2d(data);
%  % plot the data and the density estimate
%     contour3(X,Y,density,50), hold on
%     plot(data(:,1),data(:,2),'r.','MarkerSize',5)
%
% Example (Gaussian mixture with distant modes):
%
% clear all
% % generate a Gaussian mixture with distant modes
% data=[randn(100,1), randn(100,1)/4;
%       randn(100,1)+18, randn(100,1);
%       randn(100,1)+15, randn(100,1)/2-18;];
% % call the routine
%    [bandwidth,density,X,Y]=kde2d(data);
% % plot the data and the density estimate
% surf(X,Y,density,'LineStyle','none'), view([0,60])
% colormap hot, hold on, alpha(.8)
% set(gca, 'color', 'blue');
```

```
%  plot(data(:,1),data(:,2),'w.','MarkerSize',5)
%
% Example (Sinusoidal density):
%
% clear all
%  X=rand(1000,1); Y=sin(X*10*pi)+randn(size(X))/3; data=[X,Y];
% % apply routine
% [bandwidth,density,X,Y]=kde2d(data);
% % plot the data and the density estimate
% surf(X,Y,density,'LineStyle','none'), view([0,70])
% colormap hot, hold on, alpha(.8)
% set(gca, 'color', 'blue');
% plot(data(:,1),data(:,2),'w.','MarkerSize',5)
%
% Notes: If you have a more accurate density estimator
%        (as measured by which routine attains the smallest
%         L_2 distance between the estimate and the true density) or
   you have
%        problems running this code, please email me at botev@maths.
   uq.edu.au


% Reference: Botev, Z. I.,
%            "A Novel Nonparametric Density Estimator", Technical
   Report, The University of Queensland
%            http://espace.library.uq.e
%            du.au/view.php?pid=UQ:12535
global N A2 I
if nargin<2
    n=2^8;
end
n=2^ceil(log2(n)); % round up n to the next power of 2;
N=size(data,1);
if nargin<3
    MAX=max(data,[],1); MIN=min(data,[],1); Range=MAX-MIN;
    MAX_XY=MAX+Range/4; MIN_XY=MIN-Range/4;
```

```
end
scaling=MAX_XY-MIN_XY;
if N<=size(data,2)
    error('data has to be an N by 2 array where each row represents a
        two dimensional observation')
end
transformed_data=(data-repmat(MIN_XY,N,1))./repmat(scaling,N,1);
%bin the data uniformly using regular grid;
initial_data=ndhist(transformed_data,n);
% discrete cosine transform of initial data
a= dct2d(initial_data);
% now compute the optimal bandwidth^2
val=inf; t_star=0; c=0;  I=(0:n-1).^2; A2=a.^2;
while abs(val)>10^-5
    [val,t_star]=evolve(t_star);
    c=c+1; if c>10^3, error('Algorithm failed to converge in 1000
        iterations'), end
end
p_02=func([0,2],t_star);p_20=func([2,0],t_star); p_11=func([1,1],
    t_star);
t_x=(p_02^(3/4)/(4*pi*N*p_20^(3/4)*(p_11+sqrt(p_20*p_02))))^(1/3);
t_y=(p_20^(3/4)/(4*pi*N*p_02^(3/4)*(p_11+sqrt(p_20*p_02))))^(1/3);
% smooth the discrete cosine transform of initial data using t_star
a_t=exp(-(0:n-1)'.^2*pi^2*t_y/2)*exp(-(0:n-1).^2*pi^2*t_x/2).*a; %
    transpose goes with y coord.
% now apply the inverse discrete cosine transform
if nargout>1
    density=idct2d(a_t)*(numel(a_t)/prod(scaling));
    [X,Y]=meshgrid(MIN_XY(1):scaling(1)/(n-1):MAX_XY(1),MIN_XY(2):
        scaling(2)/(n-1):MAX_XY(2));
end
bandwidth=sqrt([t_x,t_y]).*scaling;
end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function   [out,time]=evolve(t)
global N
```

```
Sum_func = func([0,2],t) + func([2,0],t) + 2*func([1,1],t);
time=(2*pi*N*Sum_func)^(-1/3);
out=(t-time)/time;
end
%#########################################
function out=func(s,t)
global N
if sum(s)<=4
    Sum_func=func([s(1)+1,s(2)],t)+func([s(1),s(2)+1],t);
    time=(-2*K(s(1))*K(s(2))/N/Sum_func)^(1/(2+sum(s)));
    out=psi(s,time);
else
    out=psi(s,t);
end

end
%#########################################
function out=psi(s,Time)
global I A2
% s is a vector
w=exp(-I*pi^2*Time).*[1,.5*ones(1,length(I)-1)];
wx=w.*(I.^s(1));
wy=w.*(I.^s(2));
out=(-1)^sum(s)*(wy*A2*wx')*pi^(2*sum(s));
end
%#########################################
function out=K(s)
out=(-1)^s*prod((1:2:2*s-1))/sqrt(2*pi);
end
%#########################################
function data=dct2d(data)
% computes the 2 dimensional discrete cosine transform of data
% data is an nd cube
[nrows,ncols]= size(data);
if nrows~=ncols
    error('data is not a square array!')
```

```
end
% Compute weights to multiply DFT coefficients
w = [1;2*(exp(-i*(1:nrows-1)*pi/(2*nrows))).'];
weight=w(:,ones(1,ncols));
data=dct1d(dct1d(data)')';
    function transform1d=dct1d(x)

        % Re-order the elements of the columns of x
        x = [ x(1:2:end,:); x(end:-2:2,:) ];

        % Multiply FFT by weights:
        transform1d = real(weight.* fft(x));
    end
end
%##########################################
function data = idct2d(data)
% computes the 2 dimensional inverse discrete cosine transform
[nrows,ncols]=size(data);
% Compute wieghts
w = exp(i*(0:nrows-1)*pi/(2*nrows)).';
weights=w(:,ones(1,ncols));
data=idct1d(idct1d(data)');
    function out=idct1d(x)
        y = real(ifft(weights.*x));
        out = zeros(nrows,ncols);
        out(1:2:nrows,:) = y(1:nrows/2,:);
        out(2:2:nrows,:) = y(nrows:-1:nrows/2+1,:);
    end
end
%##########################################
function binned_data=ndhist(data,M)
% this function computes the histogram
% of an n-dimensional data set;
% 'data' is nrows by n columns
% M is the number of bins used in each dimension
% so that 'binned_data' is a hypercube with
```

```
% size length equal to M;
[nrows,ncols]=size(data);
bins=zeros(nrows,ncols);
for i=1:ncols
    [dum,bins(:,i)] = histc(data(:,i),[0:1/M:1],1);
    bins(:,i) = min(bins(:,i),M);
end
% Combine the  vectors of 1D bin counts into a grid of nD bin
% counts.
binned_data = accumarray(bins(all(bins>0,2),:),1/nrows,M(ones(1,ncols
    )));
end
```

Appendix I

Model selection

Author: Leila Muresan

```
%% Model selection
if ~qP
RR = cell2mat(res');
model = (cell2mat(BIC'));
Ttot = cell2mat(T');

MolPerClust = cell2mat(cellfun(@(list) [[1:size(list,1)]' list
    (:,4)], T, 'UniformOutput', 0)');
Lg =cellfun(@length, T);
aux  = []';
for k = 1:length(Lg)
     aux = [aux; ones(Lg(k),1)*k];
end
% 5000 if qPaint estimation failed. Ignore
id = find(MolPerClust(:,2)<5000);

% total model
[pos lk] = ModelComparison(MolPerClust(id,2)/label_eff);
%[pos lk] = ModelComparison(Ttot(:,2))
[vv modelpos] = min(lk([1:4, 6])); % Ignore log_Gaussian
disp('Model:')
switch modelpos
```

```
            case 1
              disp('Poisson')
            case 2
              disp('Exponential')
            case 3
              disp('Gaussian')
            case 4
              disp('Power Law')
            otherwise
              disp('Random')
end


AllJc = table(aux, MolPerClust(:,2)/label_eff, ones(size(aux))*
    label_eff, 'VariableNames', {'RoiNo', 'NoMolPerCluster', '
    LabelEff'});
writetable(AllJc, fullfile(resdir, strcat(strtok(fy, '.'), '
    AllJunctions.txt')));

M = table([1:size(model,1)]', model(:,1),model(:,2),model(:,3),
    model(:,4),model(:,6),...
      'VariableNames', {'ROIno','Poi','Exp', 'Normal', 'PowerLaw',
          'Random'});
totM = table( lk(:,1),lk(:,2),lk(:,3),lk(:,4),lk(:,6),...
     'VariableNames', {'Poi','Exp', 'Normal', 'PowerLaw', '
        Random'});
writetable(M, fullfile(resdir, strcat(strtok(fy, '.'), 'NegLk.
    txt')));
writetable(totM, fullfile(resdir, strcat(strtok(fy, '.'), '
    jointNegLk.txt')));
else
    data = (cell2mat(res'));
    AllJc = table((1:size(data,1))', data(:,4), 'VariableNames',
       {'RoiNo', 'NoMol'});
```

```
    writetable(AllJc, fullfile(resdir, strcat(strtok(fy, '.'), '
        Molecules.txt')));

end
```

where function **ModelComparison** is

```
function [pos lk] = ModelComparison(n)
id = find(~(isnan(n)|isinf(n)));
n = n(id);

lk = ones(1,6)*Inf;
if length(n)>2
lk(1) =log(length(n))*1 + 2*negloglik(fitdist(floor(n), 'Poisson'));
lk(2) = log(length(n))*1 + 2*negloglik(fitdist(n, 'Exponential'));
%lk(3) = negloglik(fitdist(n', 'Uniform'));
% pd = makedist('Poisson','lambda',1, 'InputData', n);
% pd.InputData = n';
% lk(3) = negloglik(pd);
%id = find(~isnan(n));

%lk(4) = negloglik(fitdist(n', 'GeneralizedPareto'));
% [alpha, xmin, lk(4)]=plfit(n')%, varargin);
% lk(4) =-lk(4);
lk(3) = log(length(n))*2 + 2*negloglik(fitdist(n, 'Normal'));
try
    lk(4) = log(length(n))*3 + 2*negloglik(fitdist(n,'
        GeneralizedPareto'));
% [alpha, xmin, lk(4)]=plfit(n')%, varargin);
% lk(4) = log(length(n))*3 - 2*lk(4);
catch
    lk(4)  = Inf;
end
lk(5) =log(length(n))*2 +2*negloglik(fitdist(n, 'LogNormal'));
lk(6) = (length(n)+sum(log(factorial(floor(n)))));
```

```
%BIC = log(length(n))*[1  1  1  2  2  2]+2*lk;

[v  pos] = min(lk);
else
    pos = 0;
end

%[aic, bic] = aicbic(logL, numParam, numObs)
```

Appendix J

Bayesian information criterion in model selection

As explained in Section 5.2.2, according to *Bayes theorem* the following holds:

$$P(M_k|D) = \frac{P(D|M_k)P(M_k)}{P(D)} \tag{J.1}$$

where $P(M_k|D)$ is the posterior probability of the model M_k given the data D, $P(D|M_k)$ is the probability to observe the data given the model M_k, $P(M_k)$ is the model prior and $P(D)$ is the probability of data.
Furthermore:

$$P(M_k|D) = \frac{P(M_k)}{P(D)} \int_{\Theta_{Mk}} p(D|M_k, \theta)p(\theta|M_k)\partial\theta, \tag{J.2}$$

where $\int_{\Theta_{Mk}} p(D|M_k, \theta)p(\theta|M_k)\partial\theta$ is called *integrated likelihood* (over all possible parameters), or *marginal likelihood* or *evidence*. Since $P(D)$ is constant and $P(M_k)$ is chosen irrespective to data, the evidence is the only term that allows the data to favour a particular model.

The evidence can be approximated by:

$$p(D_{Mk}) \sim ce^{l(\hat{\theta}}n^{-p/2} \tag{J.3}$$

for large n where n is data cardinality, $l(\hat{\theta}$ is the log-likelihood, c is a constant and p is the dimensionality of parameter space (Wit et al. 2012).

Applying a monotone decreasing transform and ignoring c, one obtains the *Bayesian information criterion*:

$$\mathrm{BIC} = -2l(\widehat{\theta}) + \ln(n)p \tag{J.4}$$

Minimizing the BIC corresponds to maximizing the posterior model probability and can be regarded as a way to select a model. Note the role of the term $ln(n)p$ to penalise complex models (with lots of parameters) and thus prevents overfitting.